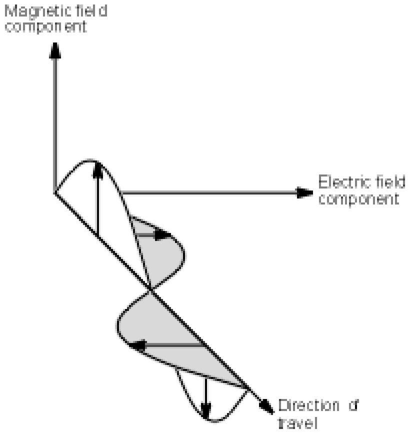
Electromagnetic waves and antenna basics

- an overview, summary, tutorial about the basics of electromagnetic waves and the way in which they affect RF antenna and RF antenna design.

Radio signals are a form of electromagnetic wave, and as they are the way in which radio signals travel, they have a major bearing on RF antennas themselves and RF antenna design.

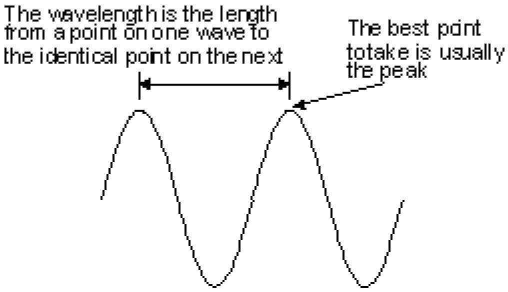
Electromagnetic waves are the same type of radiation as light, ultra-violet and infra red rays, differing from them in their wavelength and frequency. Electromagnetic waves have both electric and magnetic components that are inseparable. The planes of these fields are at right angles to one another and to the direction of motion of the wave.



An electromagnetic wave

The electric field results from the voltage changes occurring in the RF antenna which is radiating the signal, and the magnetic changes result from the current flow. It is also found that the lines of force in the electric field run along the same axis as the RF antenna, but spreading out as they move away from it. This electric field is measured in terms of the change of potential over a given distance, e.g. volts per metre, and this is known as the field strength. Similarly when an RF antenna receives a signal the magnetic changes cause a current flow, and the electric field changes cause the voltage changes on the antenna.

There are a number of properties of a wave. The first is its wavelength. This is the distance between a point on one wave to the identical point on the next. One of the most obvious points to choose is the peak as this can be easily identified although any point is acceptable.


The wavelength is the length from a point on one wave to the identical point on the next
The best point to take is usually the peak

Wavelength of an electromagnetic wave

The wavelength of an electromagnetic wave

The second property of the electromagnetic wave is its frequency. This is the number of times a particular point on the wave moves up and down in a given time (normally a second). The unit of frequency is the Hertz and it is equal to one cycle per second. This unit is named after the German scientist who discovered radio waves. The frequencies used in radio are usually very high. Accordingly the prefixes kilo, Mega, and Giga are often seen. 1 kHz is 1000 Hz, 1 MHz is a million Hertz, and 1 GHz is a thousand million Hertz i.e. 1000 MHz. Originally the unit of frequency was not given a name and cycles per second (c/s) were used. Some older books may show these units together with their prefixes: kc/s; Mc/s etc. for higher frequencies.

The third major property of the wave is its velocity. Radio waves travel at the same speed as light. For most practical purposes the speed is taken to be 300 000 000 metres per second although a more exact value is 299 792 500 metres per second.

Frequency to Wavelength Conversion

Although wavelength was used as a measure for signals, frequencies are used exclusively today. It is very easy to relate the frequency and wavelength as they are linked by the speed of light as shown:

lambda = c / f

where lambda = the wavelength in metres
f = frequency in Hertz
c = speed of radio waves (light) taken as 300 000 000 metres per second for all practical purposes.

Field measurements

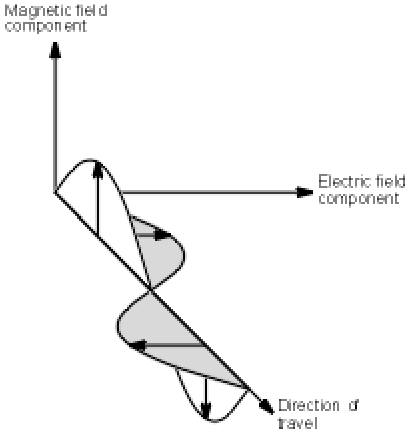
It is also interesting to note that close to the RF antenna there is also an inductive field the same as that in a transformer. This is not part of the electromagnetic wave, but it can distort measurements close to the antenna. It can also mean that transmitting antennas are more likely to cause interference when they are close to other antennas or wiring that might have the signal induced into it. For receiving antennas they are more susceptible to interference if they are close to house wiring and the like. Fortunately this inductive field falls away fairly rapidly and it is barely detectable at distances beyond about two or three wavelengths from the RF antenna.

Electromagnetic waves and antenna basics

- an overview, summary, tutorial about the basics of electromagnetic waves and the way in which they affect RF antenna and RF antenna design.

Radio signals are a form of electromagnetic wave, and as they are the way in which radio signals travel, they have a major bearing on RF antennas themselves and RF antenna design.

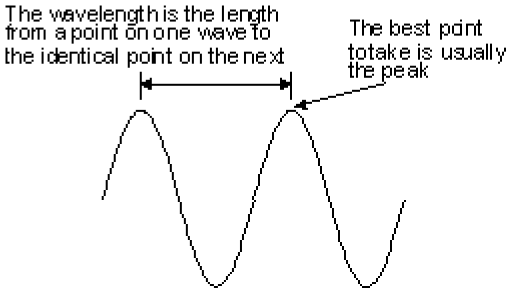
Electromagnetic waves are the same type of radiation as light, ultra-violet and infra red rays, differing from them in their wavelength and frequency. Electromagnetic waves have both electric and magnetic components that are inseparable. The planes of these fields are at right angles to one another and to the direction of motion of the wave.



An electromagnetic wave

The electric field results from the voltage changes occurring in the RF antenna which is radiating the signal, and the magnetic changes result from the current flow. It is also found that the lines of force in the electric field run along the same axis as the RF antenna, but spreading out as they move away from it. This electric field is measured in terms of the change of potential over a given distance, e.g. volts per metre, and this is known as the field strength. Similarly when an RF antenna receives a signal the magnetic changes cause a current flow, and the electric field changes cause the voltage changes on the antenna.

There are a number of properties of a wave. The first is its wavelength. This is the distance between a point on one wave to the identical point on the next. One of the most obvious points to choose is the peak as this can be easily identified although any point is acceptable.


The wavelength is the length from a point on one wave to the identical point on the next
The best point to take is usually the peak

Wavelength of an electromagnetic wave

The wavelength of an electromagnetic wave

The second property of the electromagnetic wave is its frequency. This is the number of times a particular point on the wave moves up and down in a given time (normally a second). The unit of frequency is the Hertz and it is equal to one cycle per second. This unit is named after the German scientist who discovered radio waves. The frequencies used in radio are usually very high. Accordingly the prefixes kilo, Mega, and Giga are often seen. 1 kHz is 1000 Hz, 1 MHz is a million Hertz, and 1 GHz is a thousand million Hertz i.e. 1000 MHz. Originally the unit of frequency was not given a name and cycles per second (c/s) were used. Some older books may show these units together with their prefixes: kc/s; Mc/s etc. for higher frequencies.

The third major property of the wave is its velocity. Radio waves travel at the same speed as light. For most practical purposes the speed is taken to be 300 000 000 metres per second although a more exact value is 299 792 500 metres per second.

Frequency to Wavelength Conversion

Although wavelength was used as a measure for signals, frequencies are used exclusively today. It is very easy to relate the frequency and wavelength as they are linked by the speed of light as shown:

lambda = c / f

where lambda = the wavelength in metres
f = frequency in Hertz
c = speed of radio waves (light) taken as 300 000 000 metres per second for all practical purposes.

Field measurements

It is also interesting to note that close to the RF antenna there is also an inductive field the same as that in a transformer. This is not part of the electromagnetic wave, but it can distort measurements close to the antenna. It can also mean that transmitting antennas are more likely to cause interference when they are close to other antennas or wiring that might have the signal induced into it. For receiving antennas they are more susceptible to interference if they are close to house wiring and the like. Fortunately this inductive field falls away fairly rapidly and it is barely detectable at distances beyond about two or three wavelengths from the RF antenna.

Antenna feed impedance

- overview, summary, tutorial about RF antenna or aerial feed impedance and the importance of matching RF andtennas to feeders. Radiation resistance, loss resistance, and efficiency are also detailed.

When a signal source is applied to an RF antenna at its feed point, it is found that it presents a load impedance to the source. This is known as the antenna "feed impedance" and it is a complex impedance made up from resistance, capacitance and inductance. In order to ensure the optimum efficiency for any RF antenna design it is necessary to maximise the transfer of energy by matching the feed impedance of the RF antenna design to the load. This requires some understanding of the operation of antenna design in this respect.

The feed impedance of the antenna results from a number of factors including the size and shape of the RF antenna, the frequency of operation and its environment. The impedance seen is normally complex, i.e. consisting of resistive elements as well as reactive ones.

Antenna feed impedance resistive elements

The resistive elements are made up from two constituents. These add together to form the sum of the total resistive elements.

- ***Loss resistance:*** The loss resistance arises from the actual resistance of the elements in the aRF ntenna, and power dissipated in this manner is lost as heat. Although it may appear that the "DC" resistance is low, at higher frequencies the skin effect is in evidence and only the surface areas of the conductor are used. As a result the effective resistance is higher than would be measured at DC. It is proportional to the circumference of the conductor and to the square root of the frequency.

  The resistance can become particularly significant in high current sections of an RF antenna where the effective resistance is low. Accordingly to reduce the effect of the loss resistance it is necessary to ensure the use of very low resistance conductors.
- ***Radiation resistance:*** The other resistive element of the impedance is the "radiation resistance". This can be thought of as virtual resistor. It arises from the fact that power is "dissipated" when it is radiated from the RF antenna. The aim is to "dissipate" as much

power in this way as possible. The actual value for the radiation resistance varies from one type of antenna to another, and from one design to another. It is dependent upon a variety of factors. However a typical half wave dipole operating in free space has a radiation resistance of around 73 Ohms.

Antenna reactive elements

There are also reactive elements to the feed impedance. These arise from the fact that the antenna elements act as tuned circuits that possess inductance and capacitance. At resonance where most antennas are operated the inductance and capacitance cancel one another out to leave only the resistance of the combined radiation resistance and loss resistance. However either side of resonance the feed impedance quickly becomes either inductive (if operated above the resonant frequency) or capacitive (if operated below the resonant frequency).

Efficiency

It is naturally important to ensure that the proportion of the power dissipated in the loss resistance is as low as possible, leaving the highest proportion to be dissipated in the radiation resistance as a radiated signal. The proportion of the power dissipated in the radiation resistance divided by the power applied to the antenna is the efficiency.

A variety of means can be employed to ensure that the efficiency remains as high as possible. These include the use of optimum materials for the conductors to ensure low values of resistance, large circumference conductors to ensure large surface area to overcome the skin effect, and not using designs where very high currents and low feed impedance values are present. Other constraints may require that not all these requirements can be met, but by using engineering judgement it is normally possible to obtain a suitable compromise.

Summary

It can be seen that the antenna feed impedance is particularly important when considering any RF antenna design. However by maximising the energy transfer by matching the feeder to the antenna feed impedance the antenna design can be optimised and the best performance obtained.

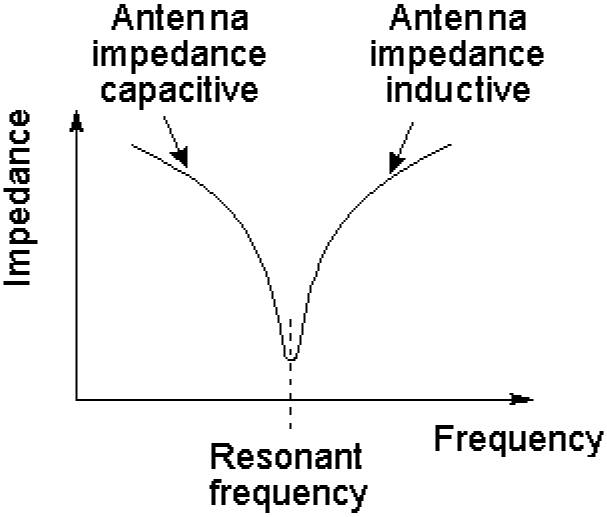
Antenna resonance and bandwidth

- overview, summary, tutorial about antenna or aerial resonance and bandwidth and the impact of RF antenna resonance and bandwidth on radio communications systems.

Two major factors associated with radio antenna design are the antenna resonant point or centre operating frequency and the antenna bandwidth or the frequency range over which the antenna design can operate. These two factors are naturally very important features of any antenna design and as such they are mentioned in specifications for particular RF ntennas. Whether the RF antenna is used for broadcasting, WLAN, cellular telecommunications, PMR or any other application, the performance of the RF antenna is paramount, and the antenna resonant frequency and the antenna bandwidth are of great importance.

Antenna resonance

An RF antenna is a form of tuned circuit consisting of inductance and capacitance, and as a result it has a resonant frequency. This is the frequency where the capacitive and inductive reactances cancel each other out. At this point the RF antenna appears purely resistive, the resistance being a combination of the loss resistance and the radiation resistance.



Impedance of an RF antenna with frequency

The capacitance and inductance of an RF antenna are determined by its physical properties and the environment where it is located. The major feature of the RF antenna design is its dimensions. It is found that the larger the antenna or more strictly the antenna elements, the lower the resonant frequency. For example antennas for UHF terrestrial television have relatively small elements, while those for VHF broadcast sound FM have larger elements indicating a lower frequency. Antennas for short wave applications are larger still.

Antenna bandwidth

Most RF antenna designs are operated around the resonant point. This means that there is only a limited bandwidth over which an RF antenna design can operate efficiently. Outside this the levels of reactance rise to levels that may be too high for satisfactory operation. Other characteristics of the antenna may also be impaired away from the centre operating frequency.

The antenna bandwidth is particularly important where radio transmitters are concerned as damage may ccur to the transmitter if the antenna is operated outside its operating range and the radio transmitter is not adequately protected. In addition to this the signal radiated by the RF antenna may be less for a number of reasons.

For receiving purposes the performance of the antenna is less critical in some respects. It can be operated outside its normal bandwidth without any fear of damage to the set. Even a random length of wire will pick up signals, and it may be possible to receive several distant stations. However for the best reception it is necessary to ensure that the performance of the RF antenna design is optimum.

Impedance bandwidth

One major feature of an RF antenna that does change with frequency is its impedance. This in turn can cause the amount of reflected power to increase. If the antenna is used for transmitting it may be that beyond a given level of reflected power damage may be caused to either the transmitter or the feeder, and this is quite likely to be a factor which limits the operating bandwidth of an antenna. Today most transmitters have some form of SWR protection circuit that prevents damage by reducing the output power to an acceptable level as the levels of reflected power increase. This in turn means that the efficiency of the station is reduced outside a given bandwidth. As far as receiving is concerned the impedance changes of the antenna are not as critical as they will mean that the signal transfer from the antenna itself to the feeder is reduced and in turn the efficiency will fall. For amateur operation the frequencies below which a maximum SWR figure of 1.5:1 is produced is often taken as the acceptable bandwidth.

In order to increase the bandwidth of an antenna there are a number of measures that can be taken. One is the use of thicker conductors. Another is the actual type of antenna used. For example a folded dipole which is described fully in Chapter 3 has a wider bandwidth than a non-folded one. In fact looking at a standard television antenna it is possible to see both of these features included.

Radiation pattern

Another feature of an antenna that changes with frequency is its radiation pattern. In the case of a beam it is particularly noticeable. In particular the front to back ratio will fall off rapidly outside a given bandwidth, and so will the gain. In an antenna such as a Yagi this is caused by a reduction in the currents in the parasitic elements as the frequency of operation is moved away from resonance. For beam antennas such as the Yagi the radiation pattern bandwidth is defined as the frequency range over which the gain of the main lobe is within 1 dB of its maximum.

For many beam antennas, especially high gain ones it will be found that the impedance bandwidth is wider than the radiation pattern bandwidth, although the two parameters are inter-related in many respects.

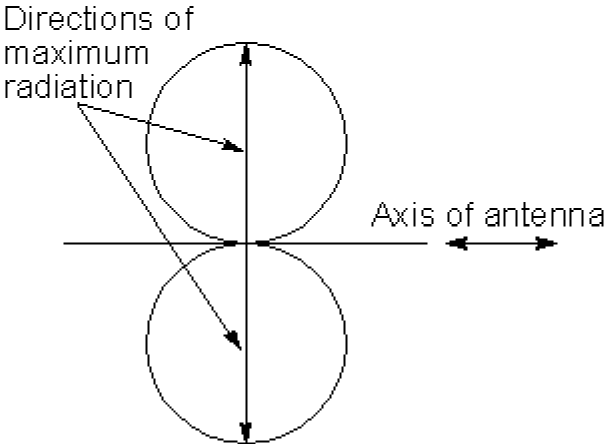
Antenna directivity and gain

- an overview, summary, tutorial about the basics of RF antenna directivity (aerial directivity) and gain including isotropic radiators, polar diagrams and antenna dBi figures and antenna dBd figures.

RF antennas or aerials do not radiate equally in all directions. It is found that any realisable RF antenna design will radiate more in some directions than others. The actual pattern is dependent upon the type of antenna design, its size, the environment and a variety of other factors. This directional pattern can be used to ensure that the power radiated is focussed in the desired directions.

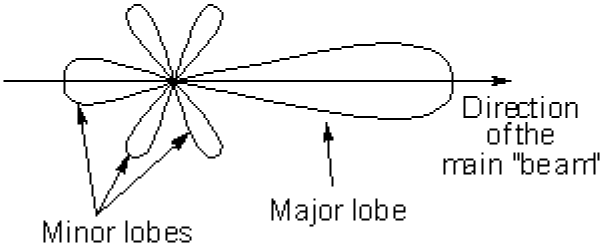
It is normal to refer to the directional patterns and gain in terms of the transmitted signal. It is often easier to visualise the RF antenna is terms of its radiated power, however the antenna performs in an exactly equivalent manner for reception, having identical figures and specifications.

In order to visualise the way in which an antenna radiates a diagram known as a polar diagram is used. This is normally a two dimensional plot around an antenna showing the intensity of the radiation at each point for a particular plane. Normally the scale that is used is logarithmic so that the differences can be conveniently seen on the plot. Although the radiation pattern of the antenna varies in three dimensions, it is normal to make a plot in a particular plane, normally either horizontal or vertical as these are the two that are most used, and it simplifies the measurements and presentation. An example for a simple dipole antenna is shown below.



Polar diagram of a half wave dipole in free space

Antenna designs are often categorised by the type of polar diagram they exhibit. For example an omni-directional antenna design is one which radiates equally (or approximately equally) in all directions in the plane of interest. An antenna design that radiates equally in all directions in all planes is called an isotropic antenna. As already mentioned it is not possible to produce one of these in reality, but it is useful as a theoretical reference for some measurements. Other RF antennas exhibit highly directional patterns and these may be utilised in a number of applications. The Yagi antenna is an example of a directive antenna and possibly it is most widely used for television reception.



Polar diagram for a yagi antenna

RF antenna beamwidth

There are a number of key features that can be seen from this polar diagram. The first is that there is a main beam or lobe and a number of minor lobes. It is often useful to define the beam-width of an RF antenna. This is taken to be angle between the two points where the power falls to half its maximum level, and as a result it is sometimes called the half power beam-width.

Antenna gain

An RF antenna radiates a given amount of power. This is the power dissipated in the radiation resistance of the RF antenna. An isotropic radiator will distribute this equally in all directions. For an antenna with a directional pattern, less power will be radiated in some directions and more in others. The fact that more power is radiated in given directions implies that it can be considered to have a gain.

The gain can be defined as a ratio of the signal transmitted in the "maximum" direction to that of a standard or reference antenna. This may sometimes be called the "forward gain". The figure that is obtained is then normally expressed in decibels (dB). In theory the standard antenna could be almost anything but two types are generally used. The most common type is a simple dipole as it is easily available and it is the basis of many other types of antenna. In this case the gain is often expressed as dBd i.e. gain expressed in decibels over a dipole. However a dipole does not radiated equally in all directions in all planes and so an isotropic source is sometimes used. In this case the gain may be specified in dBi i.e. gain in decibels over an isotropic source. The main drawback with using an isotropic source (antenna dBi) as a reference is that it is not possible to realise them in practice and so that figures using it can only be theoretical. However it is possible to relate the two gains as a dipole has a gain of 2.1 dB over an isotropic source i.e. 2.1 dBi. In other words, figures expressed as gain over an isotropic source will be 2.1 dB higher than those relative to a dipole. When choosing an antenna and looking at the gain specifications, be sure to check whether the gain is relative to a dipole or an isotropic source, i.e. the antenna dBi figure of the antenna dBd figure.

Apart from the forward gain of an antenna another parameter which is important is the front to back ratio. This is expressed in decibels and as the name implies it is the ratio of the maximum signal in the forward direction to the signal in the opposite direction. This figure is normally expressed in decibels. It is found that the design of an antenna can be adjusted to give either maximum forward gain of the optimum front to back ratio as the two do not normally coincide exactly. For most VHF and UHF operation the design is normally optimised for the optimum forward gain as this gives the maximum radiated signal in the required direction.

RF antenna gain / beamwidth balance

It may appear that maximising the gain of an antenna will optimise its performance in a system. This may not always be the case. By the very nature of gain and beamwidth, increasing the gain will result in a reduction in the beamwidth. This will make setting the direction of the antenna more critical. This may be quite acceptable in many applications, but not in others. This balance should be considered when designing and setting up a radio link.

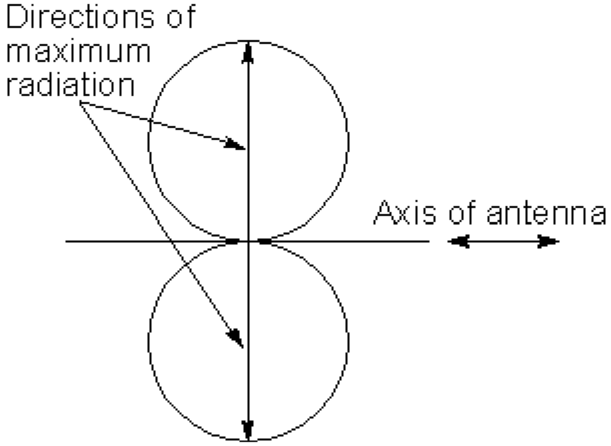
Antenna directivity and gain

- an overview, summary, tutorial about the basics of RF antenna directivity (aerial directivity) and gain including isotropic radiators, polar diagrams and antenna dBi figures and antenna dBd figures.

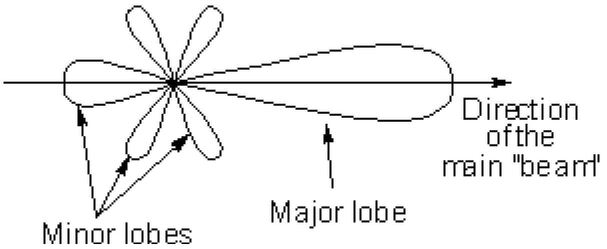
RF antennas or aerials do not radiate equally in all directions. It is found that any realisable RF antenna design will radiate more in some directions than others. The actual pattern is dependent upon the type of antenna design, its size, the environment and a variety of other factors. This directional pattern can be used to ensure that the power radiated is focussed in the desired directions.

It is normal to refer to the directional patterns and gain in terms of the transmitted signal. It is often easier to visualise the RF antenna is terms of its radiated power, however the antenna performs in an exactly equivalent manner for reception, having identical figures and specifications.

In order to visualise the way in which an antenna radiates a diagram known as a polar diagram is used. This is normally a two dimensional plot around an antenna showing the intensity of the radiation at each point for a particular plane. Normally the scale that is used is logarithmic so that the differences can be conveniently seen on the plot. Although the radiation pattern of the antenna varies in three dimensions, it is normal to make a plot in a particular plane, normally either horizontal or vertical as these are the two that are most used, and it simplifies the measurements and presentation. An example for a simple dipole antenna is shown below.



Antenna designs are often categorised by the type of polar diagram they exhibit. For example an omni-directional antenna design is one which radiates equally (or approximately equally) in all directions in the plane of interest. An antenna design that radiates equally in all directions in all planes is called an isotropic antenna. As already mentioned it is not possible to produce one of these in reality, but it is useful as a theoretical reference for some measurements. Other RF antennas exhibit highly directional patterns and these may be utilised in a number of applications. The Yagi antenna is an example of a directive antenna and possibly it is most widely used for television reception.



Polar diagram for a yagi antenna

RF antenna beamwidth

There are a number of key features that can be seen from this polar diagram. The first is that there is a main beam or lobe and a number of minor lobes. It is often useful to define the beam-width of an RF antenna. This is taken to be angle between the two points where the power falls to half its maximum level, and as a result it is sometimes called the half power beam-width.

Antenna gain

An RF antenna radiates a given amount of power. This is the power dissipated in the radiation resistance of the RF antenna. An isotropic radiator will distribute this equally in all directions. For an antenna with a directional pattern, less power will be radiated in some directions and more in others. The fact that more power is radiated in given directions implies that it can be considered to have a gain.

The gain can be defined as a ratio of the signal transmitted in the "maximum" direction to that of a standard or reference antenna. This may sometimes be called the "forward gain". The figure that is obtained is then normally expressed in decibels (dB). In theory the standard antenna could be almost anything but two types are generally used. The most common type is a simple dipole as it is easily available and it is the basis of many other types of antenna. In this case the gain is often expressed as dBd i.e. gain expressed in decibels over a dipole. However a dipole does not radiated equally in all directions in all planes and so an isotropic source is sometimes used. In this case the gain may be specified in dBi i.e. gain in decibels over an isotropic source. The main drawback with using an isotropic source (antenna dBi) as a reference is that it is not possible to realise them in practice and so that figures using it can only be theoretical. However it is possible to relate the two gains as a dipole has a gain of 2.1 dB over an isotropic source i.e. 2.1 dBi. In other words, figures expressed as gain over an isotropic source will be 2.1 dB higher than those relative to a dipole. When choosing an antenna and looking at the gain specifications, be sure to check whether the gain is relative to a dipole or an isotropic source, i.e. the antenna dBi figure of the antenna dBd figure.

Apart from the forward gain of an antenna another parameter which is important is the front to back ratio. This is expressed in decibels and as the name implies it is the ratio of the maximum signal in the forward direction to the signal in the opposite direction. This figure is normally expressed in decibels. It is found that the design of an antenna can be adjusted to give either maximum forward gain of the optimum front to back ratio as the two do not normally coincide exactly. For most VHF and UHF operation the design is normally optimised for the optimum forward gain as this gives the maximum radiated signal in the required direction.

RF antenna gain / beamwidth balance

It may appear that maximising the gain of an antenna will optimise its performance in a system. This may not always be the case. By the very nature of gain and beamwidth, increasing the gain will result in a reduction in the beamwidth. This will make setting the direction of the antenna more critical. This may be quite acceptable in many applications, but not in others. This balance should be considered when designing and setting up a radio link.

Balanced antenna feeder

- including open wire, two-wire, twin, and ribbon feeders

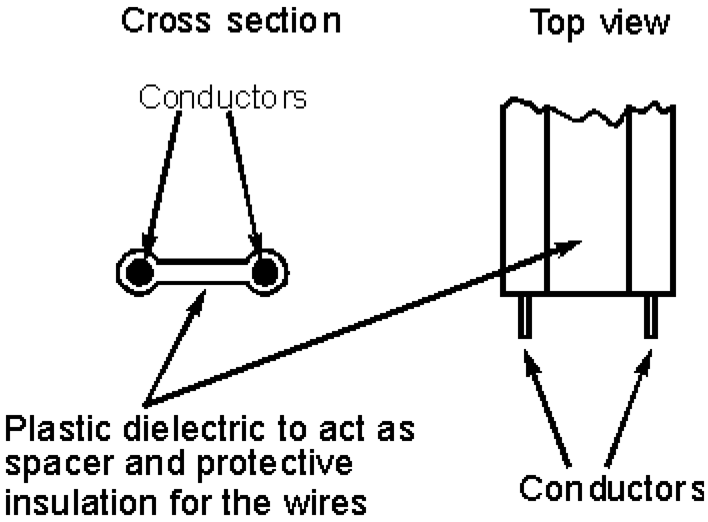
Balanced feeder is a form of feeder that can be used for feeding balanced antennas (i.e. antennas that do not have one connection taken to ground). It is mainly used on frequencies below 30 MHz can offer the advantage of very low levels of loss. The feeder or transmission line is also referred to by other names including twin, two wire, open wire, and sometimes even ribbon feeder. These names often depend upon the type of construction of the particular form.

It is used less than coaxial feeder or coax, although it is able to offer some significant advantages over coax in some applications.

Balanced feeder basics

A balanced or twin feeder consists of two parallel conductors unlike coax that consists of two concentric conductors.. The currents flowing in both wires run in opposite directions but are equal in magnitude. As a result the fields from them cancel out and no power is radiated or picked up. To ensure efficient operation the spacing of the conductors is normally kept to within about 0.01 wavelengths.

The feeder exists in a variety of forms. Essentially it is just two wires that are closely spaced in terms of the radio frequency of operation. In practical terms manufactured feeder is available and it consist of two wires contained within a plastic sheath that is also used as a spacer between them to keep the spacing, and hence the impedance constant. Another form commonly called open wire feeder simply consists of two wires kept apart by spacers that are present at regular intervals along the feeder. It has an appearance a little akin to a rope ladder.


Cross section
Top view
Conductors
Plastic dielectric to act as spacer and protective insulation for the wires
Conductors

Twin feeder a form of balanced feeder

Balanced feeder impedance

Like coaxial cable, the impedance of twin feeder is governed by the dimensions of the conductors, their spacing and the dielectric constant of the material between them. The impedance can be calculated from the formula given below.

$$Zo = \frac{276 \log(D/d)}{\sqrt{\varepsilon}}$$

Where
D is the distance between the two conductors
d is the outer diameter of the conductors
Epsilon is the dielectric constant of the material between the two conductors

Types of balanced feeder

This type of feeder can take a variety of forms. An "open wire" feeder can be made by having two wires running parallel to one another. Spacers are used every fifteen to thirty centimetres to maintain the wire spacing. Usually these are made from plastic or other insulating material. Typically this feeder may have an impedance of around 600 ohms, although it is very dependent upon the wire, and the spacing used.

The feeder may also be bought as flat 300 ohm ribbon feeder consisting of two wires spaced with a clear plastic. This is the most common form and is the type that is used for manufacturing temporary VHF FM antennas. If used outside this type absorbs water into the plastic dielectric. Not only does this significantly increase the loss on damp days, but the moisture absorbed causes the wire to oxidise which in turn leads to increased losses over the longer term.

The feeder can also be bought with a black plastic dielectric with oval holes spaced at intervals in spacing. This type gives far better performance than the clear plastic varieties which absorb water if used outside.

Coaxial feeder or RF coax cable

- an overview of coax cable often called coaxial feeder or RF cable, used to feed antennas and deliver radio frequency power from one point to another.

This coax cable tutorial is split into several pages:
[1] Coaxial feeder overview [2] Coax impedance [3] Coax cable loss or attenuation [4] Coax cable power rating [5] Coax cable velocity factor [6] Coax cable environmental factors [7] Coax cable data and specifications [8] Coaxial cable installation tips

The most common type of antenna feeder used today is undoubtedly coaxial feeder or coax cable. Coax cable, often referred to as RF cable, offers advantages of convenience of use while being able to provide a good level of performance. In view of this vast amounts of coax cable, coax feeder are manufactured each year, and it is also available in a wide variety of forms for different applications.

Applications of coax cable

Coax cable or coaxial feeder is used in many applications where it is necessary to transfer radio frequency energy from one point to another. Possibly the most obvious use of coax cable is for domestic television down-leads, but it is widely used in many other areas as well. While it is sued for domestic connections between receivers and aerials, it is likewise also used for commercial and industrial transmission lines connecting receivers and transmitters to antennas. However it is also sued where any high frequency signals need to be carried any distance. Its construction means that signals that the levels of loss and stray pick-up are minimised. In view of this it is also used in many computer applications. Coax cable was used for some early forms of Ethernet local area networks, although now optical fibres are used for higher data rates, or twisted pairs where frequencies are not so high as these cables are much cheaper than coax.

History of RF coax cable

RF coax cable is a particularly important part of today's RF and electronics scene. It is a component that could easily be overlooked with little thought of how it appeared. In the late 1800s there were a huge number of basic discoveries being made in the field of electricity. Radio, or wireless as it was originally called was not understood well, and the first transmissions were made in the 1890s. Some transmissions were made earlier but not understood.

The first known implementation of coax cable was in 1884 when Ernst von Siemens (one of the founders of the Siemens empire) patented the idea, although there were no known applications at this time. It then took until 1929 before the first modern commercial coax cables were patented by Bell Laboratories, although its use was still relatively small. Nevertheless it was used in 1934 to relay television pictures of the Berlin Olympics to Leipzig. Then in 1936 an a coaxial cable was installed between London and Birmingham in the UK to carry 40 telephone calls, and in the USA an experimental coaxial cable was installed between New York and Philadelphia to relay television pictures.

With the commercial use of RF coax cable establishing itself, many other used the cable for shorter runs. It quickly established itself, and now it is widely used for both commercial and domestic applications.

Coax cable basics

Coax cable, coaxial feeder is normally seen as a thick electrical cable. The cable is made from a number of different elements that when together enable the coax cable to carry the radio frequency signals with a low level of loss from one location to another. The main elements within a coax cable are:

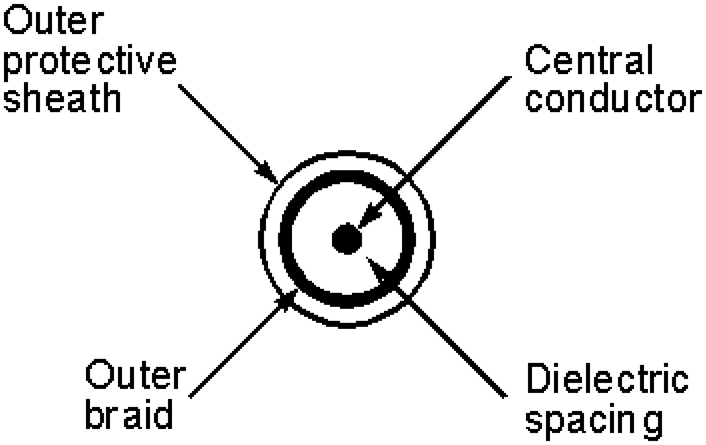
1. Centre conductor

2. Insulating dielectric

3. Outer conductor

4. Outer protecting jacket or sheath

The overall construction of the coax cable or RF cable can be seen in the diagram below and from this it can be seen that it is built up from a number of concentric layers. Although there are many varieties of coax cable, the basic overall construction remains the same:



Cross section though coaxial cable

1. ***Centre conductor*** The centre conductor of the coax is almost universally made of copper. Sometimes it may be a single conductor whilst in other RF cables it may consist of several strands.

2. ***Insulating dielectric*** Between the two conductors of the coax cable there is an insulating dielectric. This holds the two conductors apart and in an ideal world would not introduce any loss, although it is one of the chief causes of loss in reality. This coax cable dielectric may be solid or as in the case of many low loss cables it may be semi-airspaced because it is the dielectric that introduces most of the loss. This may be in the form of long "tubes" in the dielectric, or a "foam" construction where air forms a major part of the material.

3. ***Outer conductor*** The outer conductor of the RF cable is normally made from a copper braid. This enables the coax cable to be flexible which would not be the case if the outer conductor was solid, although in some varieties made for particular applications it is. To improve the screening double or even triple screened coax cables are sometimes used. Normally this is accomplished by placing one braid directly over another although in some instances a copper foil or tape outer may be used. By using additional layers of screening, the levels of stray pick-up and radiation are considerably reduced. The loss is marginally lower.

4. ***Outer protecting jacket or sheath*** Finally there is a final cover or outer sheath to the coax cable. This serves little electrical function, but can prevent earth loops forming. It also gives a vital protection needed to prevent dirt and moisture attacking the cable, and prevent the coax cable from being damaged by other mechanical means.

How RF coax cable works

A coaxial cable carries current in both the inner and the outer conductors. These current are equal and opposite and as a result all the fields are confined within the cable and it neither radiates nor picks up signals.

This means that the cable operates by propagating an electromagnetic wave inside the cable. As there are no fields outside the coax cable it is not affected by nearby objects. Accordingly it is ideal for applications where the RF cable has to be routed through or around buildings or close to many other objects. This is a particular advantage of coaxial feeder when compared with other forms of feeder such as two wire (open wire, or twin) feeder.

- **Home**
-
- Antennas & propagation
- Broadcast technology
- Cellular telecoms
- Circuits
- Design principles
- Electronic components
- Electronics manufacture
- Formulae & data
- Radio receivers
- RF topics
- Satellite technology
- Technology history
- Telecoms & networks
- Test & measurement

RF Coax Cable Connectors

- an overview of the different types of RF connectors that can be used with coax cable or feeder including the UHF (SO239 / PL259), BNC, TNC, N-type, SMA, SMB, SMC, MCX, etc..

This RF coax cable conenctors reference is split into several pages each of which addresses different aspects of RF coax cable connector operation and technology:

[1] RF coax cable connectors overview [2] BNC connector [3] TNC connector [4] C-type connector [5] N-type connector [6] SMA connector [7] SMB connector [8] UHF connector, SO239 and PL259

Coax cable connectors, often called RF connectors are in widespread use. Wherever radio frequency or RF connections need to be made there is the possibility of using coaxial connectors. Where signals reach frequencies above a few million Hertz, these coaxial connectors need to be used. The need for their use arises because it is necessary to transfer radio frequency, RF, energy from one place to another using a transmission line. The most convenient, and hence the most commonly used form of transmission line is coaxial cable which consists of two concentric conductors, an inner conductor and an outer conductor, often called the screen. Between these two conductors there is an insulating dielectric.

Coaxial cable has a number of properties, one of which is the characteristic impedance. In order that the maximum power transfer takes place from the source to the load, the characteristic impedances of both should match. Thus the characteristic impedance of a feeder is of great importance. Any mismatch will result in power being reflected back towards the source.

It is also important that RF coaxial cable connectors have a characteristic impedance that matches that of the cable. If not, a discontinuity is introduced and losses may result.

There is a variety of connectors that are used for RF applications. Impedance, frequency range, power handling, physical size and a number of other parameters including cost will determine the best type for a given applications.

UHF connector

The UHF connector, also sometimes known as the Amphenol coaxial connector was designed in the 1930s by a designer in the Amphenol company for use in the radio industry. The plug may be referred to as a PL259 coaxial connector, and the socket as an SO239 connector. These are their original military part numbers

These coaxial connectors have a threaded coupling, and this prevents them from being removed accidentally. It also enables them to be tightened sufficiently to enable a good low resistance connection to be made between the two halves.

The drawback of the UHF or Amphenol connector is that it has a non-constant impedance. This limits their use to frequencies of up to 300 MHz, but despite this these UHF connectors provide a low cost connector suitable for many applications, provided that the frequencies do not rise. Also very low cost versions are available for applications such as CB operation, and these are not suitable for operation much above 30 MHz. In view of their non-constant impedance, these connectors are now rarely used for many professional applications, being generally limited to CB, amateur radio and some video and public address systems.

N-type connector

The N-type connector is a high performance RF coaxial connector used in many RF applications. This coax connector was designed by Paul Neill of Bell Laboratories, and it gained its name from the first letter of his surname.

This RF connector has a threaded coupling interface to ensure that it mates correctly. It is available in either 50 ohm or 75 ohm versions. These two versions have subtle mechanical differences that do not allow the two types to mate. The connector is able to withstand relatively high powers when compared to the BNC or TNC connectors. The standard versions are specified for operation up to 11 GHz, although precision versions are available for operation to 18GHz.

The N-type coaxial connector is used for many radio frequency applications including broadcast and communications equipment where its power handling capability enables it to be used for medium power transmitters, however it is also used for many receivers and general RF applications.

BNC connector

The BNC coax connector is widely used in professional circles being used on most oscilloscopes and many other laboratory instruments. The BNC connector is also widely used when RF connections need to be made. The BNC connector has a bayonet fixing to prevent accidental disconnection while being easy to disconnect when necessary. This RF connector was developed in the late 1940s and it gains its name from a combination of the fact that it has a bayonet fixing and from the names of the designers, the letters BNC standing for **B**ayonet **N**eill **C**oncelman. In fact the BNC connector is essentially a miniature version of the C connector which was a bayonet version of the N-type connector.

Electrically the BNC coax cable connector is designed to present a constant impedance and it is most common in its 50 ohm version, although 75 ohm ones can be obtained. It is recommended for operation at frequencies up to 4 GHz and it can be used up to 10 GHz provided the special top quality versions specified to that frequency are used.

TNC connector

The TNC connector is very similar to the BNC connector. The main difference is that it has a screw fitting instead of the bayonet one. The TNC connector was developed originally to overcome problems during vibration. As the bayonet fixing moved slightly there were small changes to the resistance of the connections and this introduced noise. To solve the problem a screw fixing was used and the TNC coax cable connector gains its name from the words **T**hreaded **N**eill **C**oncelman.

Like the BNC connector, the TNC connector has a constant impedance, and in view of the threaded connection, its frequency limit can be extended. Most TNC connectors are specified to 11 GHz, and some may be able to operate to 18 GHz.

SMA connector

This sub-miniature RF coaxial cable connector takes its name from the words **S**ub-**M**iniature **A** connector. It finds many applications for providing connectivity for RF assemblies within equipments. It is often used for providing RF connectivity between boards, and many microwave components including filters, attenuators, mixers and oscillators, use SMA connectors.

The connectors have a threaded outer coupling interface that has a hexagonal shape, allowing it to be tightened with a spanner. Special torque spanners are available to enable them to be tightened to the correct tightness, allowing a good connection to be made without over-tightening them.

The SMA connector was originally designed in the 1960s for use with 141 semi-rigid coax cable. Here the centre of the coax forms the centre pin for the connection, removing the necessity for a transition between the coax centre conductor and a special connector centre pin. However its use extended to other flexible cables, and connectors with centre pins were introduced.

SMA connectors are regularly used for frequencies well into the microwave region, and some versions may be used at frequencies up to 26.5 GHz. For flexible cables, the frequency limit is normally determined by the cable and not the connector.

SMB connector

The SMB connector derives its name as it is termed a **S**ub-**M**iniature **B** connector. It was developed as a result of the need for a connector that was able to connect and disconnect swiftly. It does not require nuts to be tightened when two connectors are mated. Instead the connectors are brought together and they snap fit together. Additionally the connector utilizes an inner contact and overlapping dielectric insulator structures to ensure good connectivity and a constant impedance.

SMB coaxial connectors perform well under moderate vibration only and the 50 ohm versions are often specified to 4 GHz. 75 ohm versions of the SMB coaxial connector are also available, but there are often not specified up to the same frequencies, often only about 2GHz.

SMB coaxial connectors are not as widely used as their SMA counterparts. They are used for inter board or assembly connections within equipment, although they are not widely used for purchased microwave assemblies in view of their inferior performance.

SMC connector

A third SM type connector is not surprisingly the **S**ub **M**iniature **C** or SMC coaxial cable connector. It is similar to the SMB connector, but it uses a threaded coupling interface rather than the snap-on connection. This provides a far superior interface for the connection and as a result, SMC coaxial cable connectors are normally specified to operate at frequencies up to 10 GHz.

SMC coaxial cable connectors provide a good combination of small size and performance. They may also be used in environments where vibration is anticipated. In view of their performance they find applications in microwave equipment, although they are not normally used for military applications where SMA connectors tend to be preferred.

MCX connector

A number of mico-miniatiure RF connectors have been developed by a variety of manufacturers to meet the growing demand for cost effective, high quality smaller connectors. These are finding high levels of use, for example in the cellular phone industry, where size, cost and performance are all important. In fact the MCX is about 30% smaller in both size and weight than an SMB connector to which it has many similarities.

One connector that falls into this category is the MCX (**M**icro**C**oa**X**) coax connector. This was developed in the 1980s by Huber and Suhner of which MCX is a trade name. The MCX connector has many similarities with the construction of the SMB connector using a quick snap-on interface, and utilising an inner contact and an overlapping dielectric insulator structure.

The MCX connector is normally specified for operation up to 6 GHz, and it finds applications in a variety of arenas including equipment for cellular telecommunications, data telemetry, Global positioning (GPS) and other applications where size and weight are important and frequencies are generally below 5 GHz.

MMCX connector

Another connector which is being widely used is the MMCX connector. Being some 45% smaller than an SMB connector, the MMCX is ideal where a low profile outline is a key element. It is therefore ideal for applications where board height is limited, including applications where boards may be stacked. As such it is being widely used in many cellular telecommunications applications.

The connector provides a snap fitting and also utilises a slot-less design to minimise leakage.

Overview

There is a great variety of RF coaxial cable connectors in use today. The list above describes some of the more popular types of RF connector, but there are nevertheless more varieties available. When choosing a coaxial cable connector, the requirements should be carefully matched to the available options to see which RF connector will provide the best choice. In this way the best compromise between size, weight, performance and cost can be achieved.

Waveguide basics tutorial

- overview, summary, tutorial about the basics of what is a waveguide and the basic waveguide theory.

This waveguide tutorial is split into several pages:

[1] Waveguide basics [2] Waveguide theory [3] Waveguide impedance [4] Flexible waveguide [5] Waveguide couplers and transitions [6] Waveguide dimensions and sizes [7] Waveguide flanges

RF waveguides are a form of RF feeder used for microwave applications. For use, the basics of waveguide technology are easy to comprehend, although the mathematics involved can become complicated when wave theory and the like are used. Waveguides are a guide for electromagnetic waves and this gives rise to the name waveguide.

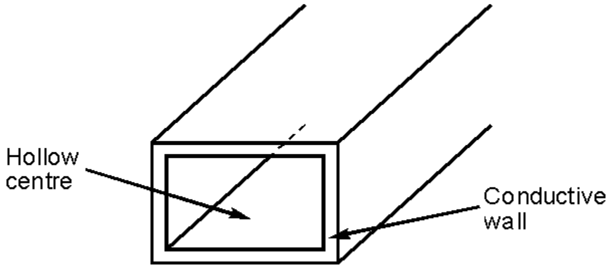
Waveguides effectively confirm and direct a wave within a given boundary. Within many electronic circles, waveguides are most commonly used for microwave RF signals, the same principles can be used for many forms of wave from sound to light.

RF waveguide basics

Waveguides are used in a variety of applications to carry radio frequency energy from one pint to another. In their broadest terms a waveguide is described as a system of material that is designed to confine electromagnetic waves in a direction defined by its physical boundaries. This definition gives a very broad view of waveguides, but indicates that waveguide theory can be applied in a number of areas and in a variety of different ways.

Electromagnetic waves propagating in open space travel out in all directions and can be thought of as spherical waves travelling out from a central source. As a result the power intensity decreases as the distance increases - it is proportional to the power of the source divided by the square of the distance. The waveguide operates by confining the electromagnetic wave so that it does not spread out and losses resulting from this effect are eliminated.

Typically a waveguide is thought if as a transmission line comprising a hollow conducting tube, which may be rectangular or circular within which electromagnetic waves are propagated. Unlike coaxial cable, there is no centre conductor within the waveguide. Signals propagate within the confines of the metallic walls that act as boundaries. The signal is confined by total internal reflection from the walls of the waveguide.



Rectangular waveguide

Waveguides will only carry or propagate signals above a certain frequency, known as the cut-off frequency. Below this the waveguide is not able to carry the signals. The cut-off frequency of the waveguide depends upon its dimensions. In view of the mechanical constraints this means that waveguides are only used for microwave frequencies. Although it is theoretically possible to build waveguides for lower frequencies the size would not make them viable to contain within normal dimensions and their cost would be prohibitive.

As a very rough guide to the dimensions required for a waveguide, the width of a waveguide needs to be of the same order of magnitude as the wavelength of the signal being carried. As a result, there is a number of standard sizes used for waveguides as detailed in another page of this tutorial. Also other forms of waveguide may be specifically designed to operate on a given band of frequencies

Types of RF waveguide

There is a number of different types of RF waveguide that can be used, bought and designed.

Typically waveguides are thought of as being rectangular in cross section as this is the most common form of waveguide. However other types and approaches may be used.

- ***Rectangular waveguide:*** This is the most commonly used form of waveguide and has a rectangular cross section.
- ***Circular waveguide:*** Circular waveguide is less common than rectangular waveguide. They have many similarities in their basic approach, although signals often use a different mode of propagation.
- ***Circuit board stripline:*** This form of waveguide is used on printed circuit boards as a transmission line for microwave signals. It typically consists of a line of a given thickness above an earth plane. Its thickness defines the impedance.

In addition to these basic forms, there are also flexible waveguides. These are most widely seen in the rectangular format. Flexible waveguide is often used to connect to antennas, etc that may not be fixed or may be moveable.

Summary

Waveguides are more expensive than other forms of RF feeder. However waveguides offer a number of advantages that mean they are the only feasible solution in many applications. Although waveguide is not nearly as widely used as other forms of feeder such as coax, it still forms and essential method of transferring RF power, especially are microwave frequencies.

Dipole antenna

- overview, summary, tutorial about the basics of the dipole antenna or dipole aerial that is widely used on its own and as the basis for other RF antenna designs.

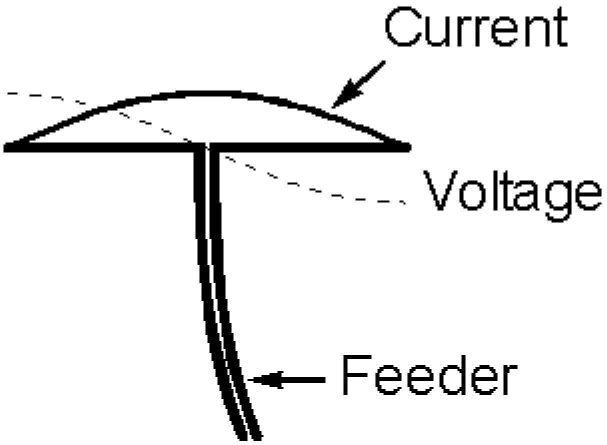
This dipole antenna tutorial is split into several pages each of which addresses a different aspect of dipole antenna technology:

[1] Dipole antenna [2] half wave dipole antenna

The dipole antenna or dipole aerial is one of the most important and commonly used types of RF antenna. It is widely used on its own, and it is also incorporated into many other RF antenna designs where it forms the radiating or driven element for the antenna.

Dipole antenna basics

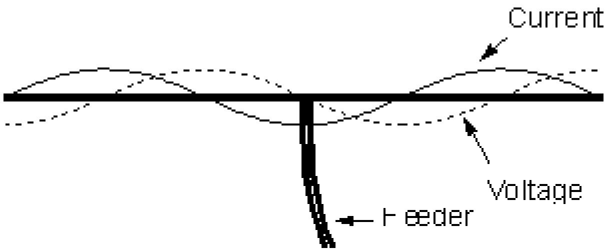
As the name suggests the dipole antenna consists of two terminals or "poles" into which radio frequency current flows. This current and the associated voltage causes and electromagnetic or radio signal to be radiated. Being more specific, a dipole is generally taken to be an antenna that consists of a resonant length of conductor cut to enable it to be connected to the feeder. For resonance the conductor is an odd number of half wavelengths long. In most cases a single half wavelength is used, although three, five, �. wavelength antennas are equally valid.



The basic half wave dipole antenna

The current distribution along a dipole is roughly sinusoidal. It falls to zero at the end and is at a maximum in the middle. Conversely the voltage is low at the middle and rises to a maximum at the ends. It is generally fed at the centre, at the point where the current is at a maximum and the voltage a minimum. This provides a low impedance feed point which is convenient to handle. High voltage feed points are far less convenient and more difficult to use.

When multiple half wavelength dipoles are used, they are similarly normally fed in the centre. Here again the voltage is at a minimum and the current at a maximum. Theoretically any of the current maximum nodes could be used.



Three half wavelength wave dipole antenna

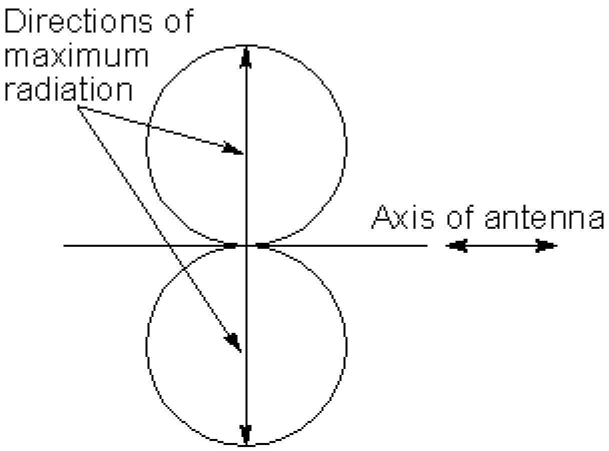
Dipole feed impedance

As with any RF antenna, the feed impedance of a dipole antenna is dependent upon a variety of factors including the length, the feed position, the environment and the like. A half wave centre fed dipole antenna in free space has an impedance 73.13 ohms making it ideal to feed with 75 ohm feeder.

The feed impedance of a dipole can be changed by a variety of factors, the proximity of other objects having a marked effect. The ground has a major effect. If the dipole antenna forms the radiating element for a more complicated form of RF antenna, then elements of the RF antenna will have an effect. Often the effect is to lower the impedance, and when used in some antennas the feed impedance of the dipole element may fall to ten ohms or less, and methods need to be used to ensure a good match is maintained with the feeder.

Dipole polar diagram

The polar diagram of a half wave dipole antenna that the direction of maximum sensitivity or radiation is at right angles to the axis of the RF antenna. The radiation falls to zero along the axis of the RF antenna as might be expected.



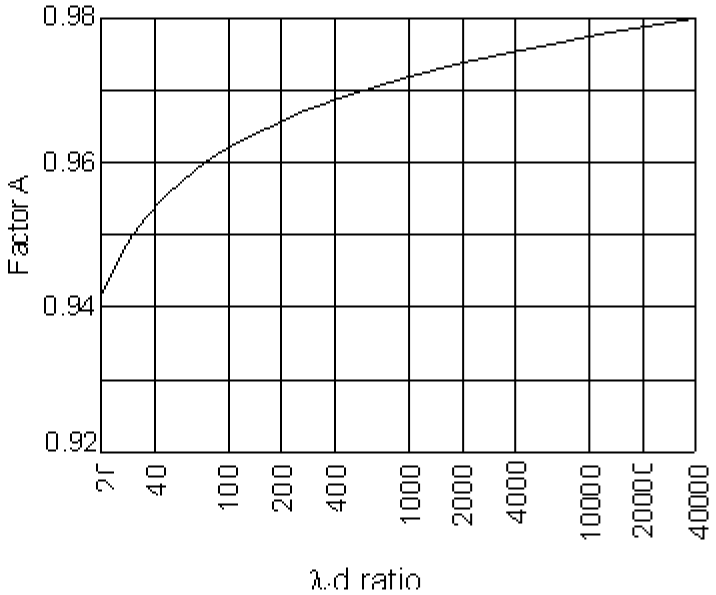
Polar diagram of a half wave dipole in free space

If the length of the dipole antenna is changed then the radiation pattern is altered. As the length of the antenna is extended it can be seen that the familiar figure of eight pattern changes to give main lobes and a few side lobes. The main lobes move progressively towards the axis of the antenna as the length increases.

Antenna length

The length of a dipole is the main determining factor for the operating frequency of the dipole antenna. Although the antenna may be an electrical half wavelength, or multiple of half wavelengths, it is not exactly the same length as the wavelength for a signal travelling in free space. There are a number of reasons for this and it means that an antenna will be slightly shorter than the length calculated for a wave travelling in free space.

For a half wave dipole the length for a wave travelling in free space is calculated and this is multiplied by a factor "A". Typically it is between 0.96 and 0.98 and is mainly dependent upon the ratio of the length of the antenna to the thickness of the wire or tube used as the element. Its value can be approximated from the graph:



Multiplication factor "A" used for calculating the length of a dipole

In order to calculate the length of a half wave dipole the simple formulae given below can be used:

Length (metres) = 150 x A / frequency in MHz

Length (inches) = 5905 x A / frequency in MHz

Using these formulae it is possible to calculate the length of a half wave dipole. Even though calculated lengths are normally quite repeatable it is always best to make any prototype antenna slightly longer than the calculations might indicate. This needs to be done because changes in the thickness of wire being used etc may alter the length slightly and it is better to make it slightly too long than too short so that it can be trimmed so that it resonates on the right frequency. It is best to trim the antenna length in small steps because the wire or tube cannot be replaced very easily once it has been removed.

Dipole basics summary

The dipole antenna is a particualrly important form of RF antenna which is very widely used for radio transmitting and receiving applications. The dipole is often used on its own as an RF antenna, but it also forms the essential element in many other types of RF antenna. As such it is the possibly the most important form of RF antenna.

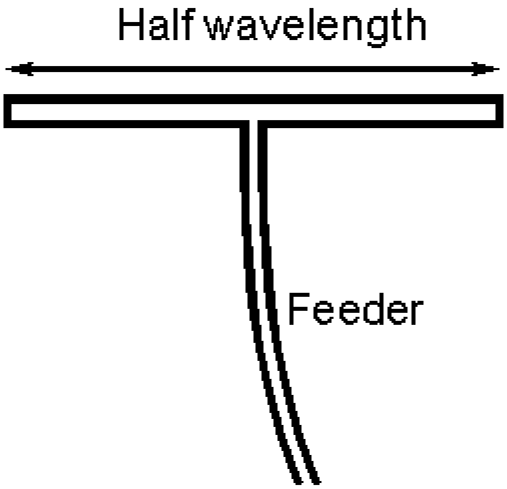
This dipole antenna tutorial is split into several pages each of which addresses a different aspect of dipole antenna technology:

[1] Dipole antenna [2] half wave dipole antenna

The standard dipole is widely used in its basic form. However under a number of circumstances a modification of the basic dipole, known as a folded dipole provides a number of advantages that can be used to advantage.

In its basic form a dipole consists of a single wire or conductor cut in the middle to accommodate the feeder. It is found that the feed impedance is altered by the proximity of other objects, especially other parasitic elements that may be used in other forms of RF antenna design. This can cause problems with matching and because resistance losses in the RF antenna system can start to become significant.

Additionally many RF antennas have to be able to operate over large bandwidths and a standard dipole may be unable to fulfil this requirement adequately.



The basic folded half wave dipole

A variation of the dipole, known as a folded dipole provides a solution to these problems, offering a wider bandwidth and a considerable increase in feed impedance. The folded dipole is formed by taking a standard dipole and then taking a second conductor and joining the two ends. In this way a complete loop is made as shown. If the conductors in the main dipole and the second or "fold" conductor are the same diameter, then it is found that there is a fourfold increase in the feed impedance. In free space, this gives a feed impedance of around 300 ohms. Additionally the RF antenna has a wider bandwidth.

Impedance increase

In a standard dipole the currents flowing along the conductors are in phase and as a result there is no cancellation of the fields and radiation occurs. When the second conductor is added this can be considered as an extension to the standard dipole with the ends folded back to meet each other. As a result the currents in the new section flow in the same direction as those in the original dipole. The currents along both the half-waves are therefore in phase and the antenna will radiate with the same radiation patterns etc as a simple half-wave dipole.

The impedance increase can be deduced from the fact that the power supplied to a folded dipole is evenly shared between the two sections which make up the antenna. This means that when compared to a standard dipole the current in each conductor is reduced to a half. As the same power is applied, the impedance has to be raised by a factor of four to retain balance in the equation Watts = I^2 x R.

Folded dipole applications

Folded dipoles are sometimes used on their own, but they must be fed with a high impedance feeder, typically 300 ohms. However they find more uses when a dipole is incorporated in another RF antenna design with other elements nearby. This has the effect of reducing the dipole impedance. To ensure that it can be fed conveniently, a folded dipole may be used to raise the impedance again to a suitable value.

Distributed Antenna System DAS

- overview, summary, tutorial about the DAS, Distributed Antenna Systems technology used for gaining better coverage and using a lower power.

The concept of a Distributed Antenna System, DAS has many advantages in some applications. A Distributed antenna system, DAS is a network of antennas spaced apart from each other, but connected to a common source. In this way the DAS is able to provide wireless or radio coverage within a given area.

The idea of a distributed antenna system is being adopted increasingly as it enables a number of advantages to be gained. However this is at the cost of a larger more complicated system. Nevertheless, distributed antenna systems are being used in a variety of areas to enable the right coverage to be gained for several applications.

Although the concept of distributed antenna systems has been known about for many years, it is with the increased deployment of wireless systems within buildings and other difficult coverage areas that the idea of distributed antenna systems has come to the fore.

Distributed antenna system advantages and disadvantages

Advantages of using a distributed antenna system

- Better defined coverage
- Fewer coverage holes
- Same coverage using a lower overall power
- Lowers health risk as a result of using lower overall power levels
- Individual antennas do not need to be as high as a single antenna for the equivalent coverage

Disadvantages of using a distributed antenna system

- Higher cost as a result of additional infrastructure required
- Possible greater visual impact in some applications as a result of greater number of antennas, although they are likely to be much lower in height.

Basic concept of a distributed antenna system

The basic idea behind the distributed antenna system is to utilise several different antennas over the required coverage area. Using this approach the overall power required is less because these more localised antennas can be placed more effectively for a small area, rather than having a single, larger antenna that is a compromise for the wider coverage needed. By adopting a distributed antenna system approach, this helps overcome the shadowing and penetration losses because a line of sight link is available more frequently. As a result the levels of absorption are lower and this means the overall power levels can be reduced.

The Yagi antenna

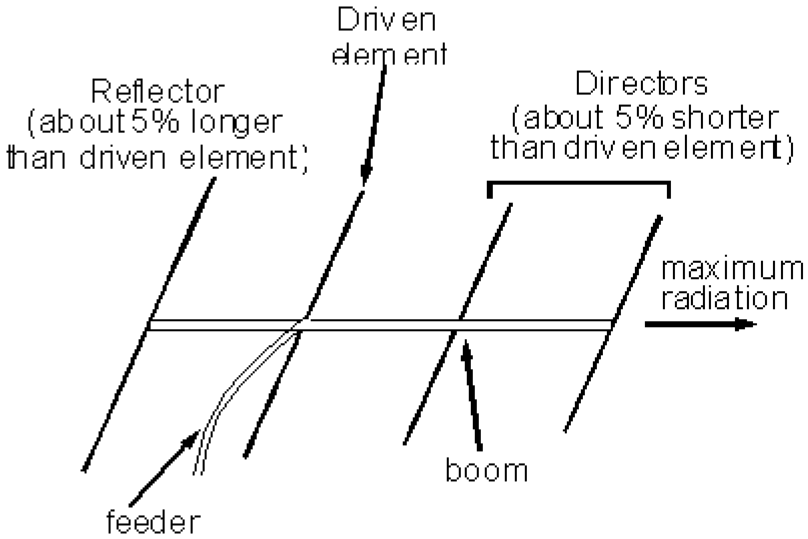
- overview, summary, tutorial about the Yagi antenna sometimes called the Yagi-Uda RF antenna that is widely used where gain and directivity are required from an RF antenna design.

The Yagi or Yagi-Uda RF antenna or aerial is one of the most successful RF antenna designs for directive applications. It is used in a wide variety of applications where an RF antenna design with gain and directivity is required. It has become particularly popular for television reception, but it is used in very many other applications where an RF antenna design is needed that has gain.

The full name for the antenna is the Yagi-Uda antenna. It was derives it name from its two Japanese inventors Yagi and his student Uda. The RF antenna design concept was first outlined in a paper that Yagi himself presented in 1928. Since then its use has grown rapidly to the stage where today a television antenna is synonymous with an RF antenna having a central boom with lots of elements attached.

The Yagi antenna

The Yagi RF antenna design has a dipole as the main radiating or driven element. Further "parasitic" elements are added which are not directly connected to the driven element. Instead they pick up power from the dipole and re-radiate it such a manner that it affects the properties of the RF antenna as a whole.


Driven element
Reflector (about 5% longer than driven element)
Directors (about 5% shorter than driven element)
maximum radiation
boom
feeder

Basic concept of a Yagi antenna

The parasitic elements of the Yagi antenna operate by re-radiating their signals in a slightly different phase to that of the driven element. In this way the signal is reinforced in some directions and cancelled out in others. It is found that the amplitude and phase of the current that is induced in the parasitic elements is dependent upon their length and the spacing between them and the dipole or driven element.

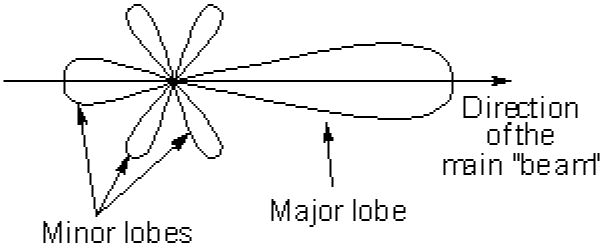
Using a parasitic element it is not possible to have complete control over both the amplitude and phase of the currents in all the elements. This means that it is not possible to obtain complete cancellation in one direction. Nevertheless it is still possible to obtain a high degree of reinforcement in one direction and have a high level of gain, and also have a high degree of cancellation in another to provide a good front to back ratio.

To obtain the required phase shift an element can be made either inductive or capacitive. If the parasitic element is made inductive it is found that the induced currents are in such a phase that they reflect the power away from the parasitic element. This causes the RF antenna to radiate more power away from it. An element that does this is called a reflector. It can be made inductive by tuning it below resonance. This can be done by physically adding some inductance to the element in the form of a coil, or more commonly by making it longer than the resonant length. Generally it is made about 5% longer than the driven element.

If the parasitic element is made capacitive it will be found that the induced currents are in such a phase that they direct the power radiated by the whole antenna in the direction of the parasitic element. An element which does this is called a director. It can be made capacitive tuning it above resonance. This can be done by physically adding some capacitance to the element in the form of a capacitor, or more commonly by making it about 5% shorter than the driven element.

It is found that the addition of further directors increases the directivity of the antenna, increasing the gain and reducing the beamwidth. The addition of further reflectors makes no noticeable difference.

The antenna exhibits a directional pattern consisting of a main forward lobe and a number of spurious side lobes. The main one of these is the reverse lobe caused by radiation in the direction of the reflector. The antenna can be optimised to either reduce this or produce the maximum level of forward gain. Unfortunately the two do not coincide exactly and a compromise on the performance has to be made depending upon the application.



Polar diagram of the Yagi antenna

Gain

The gain of a Yagi antenna is governed mainly by the number of elements in the particular RF antenna. However the spacing between the elements also has an effect. As the overall performance of the RF antenna has so many inter-related variables, many early designs were not able to realise their full performance. Today computer programmes are used to optimise RF antenna designs before they are even manufactured and as a result the performance of antennas has been improved.

Feed impedance

It is possible to vary the feed impedance of a Yagi antenna over a wide range. Although the impedance of the dipole itself would be 73 ohms in free space, this is altered considerably by the proximity of the parasitic elements. The spacing, their length and a variety of other factors all affect the feed impedance presented by the dipole to the feeder. In fact altering the element spacing has a greater effect on the impedance than it does the gain, and accordingly setting the required spacing can be used as one design technique to fine tune the required feed impedance. Nevertheless the proximity of the parasitic elements usually reduces the impedance below the 50 ohm level normally required. It is found that for element spacing distances less than 0.2 wavelengths the impedance falls rapidly away.

To overcome this, a variety of techniques can be used. One is to use a folded dipole for the driven element. This provides an increase in impedance of around four times dependent upon the ratio of the thicknesses of the basic dipole conductor and the "fold" conductor. Other techniques involve using gamma matches, delta matches, baluns and the like. Delta matches can be very convenient. They involve "fanning out" the connection to the driven element. This method has the advantage that the driven element does not need to be broken to apply the feed as shown. As this is really applicable to a balanced feeder, a balun is required if coaxial cable is to be used.

A gamma match is another alternative that is often used. The outer or braid of the coax feeder is connected directly to the centre of the driven element. This can be done because the RF voltage at the centre is zero at this point. The inner conductor of the feeder carrying the RF current is taken out along the driven element. The inductance of the arm is then tuned out by the variable capacitor. When adjusting the RF antenna design, both the variable capacitor and the point at which the arm contacts the driven element are adjusted. Once a value has been ascertained for the variable capacitor, its value can be measured and a fixed component inserted if required.

Summary

The Yagi antenna is a particulary useful form of RF antenna design. It is widely used in applications where an RF antenna design is required to provide gain and directivity. In this way the optimum transmission and reception conditions can be obtained.

Satellite antenna

- Overview, of the typical RF antenna design types used with satellites, both on the ground and on the satellite. This includes satellite television (tv) reception.

A variety of forms of antenna can be used for transmitting to and receiving from satellites. The most common type of satellite antenna is the parabolic reflector, however this is not the only type of antenna that can be used. The actual type of antenna will depend upon what the overall application and the requirements.

Antenna gain

The distances over which signals travel to some satellites is very large. Geostationary ones are a particular case. This means that path losses are high and accordingly signal levels are low. In addition to this the power levels that can be transmitted by satellites are limited by the fact that all the power has be generated from solar panels. As a result the antennas that are used are often high gain directional varieties. The parabolic reflector is one of the most popular.

Antennas on satellites

Although there is fundamentally no difference between the antennas on satellites and those on the ground there are a number of different requirements that need to be taken into account. In the first instance the environmental conditions are very different. As conditions in space are particularly harsh the antennas need to be built to withstand this. Temperatures vary considerably between light and dark and this will cause expansion and contraction. The materials that are sued in the conduction need to be carefully chosen.

The gain and directivity of the antenna need to be chosen to meet the needs of the satellite. For most geostationary satellites the use of directional antennas with gain is mandatory in view of the path losses incurred. These satellites are more likely to cover a give area of the Earth, and as they remain in the same position this is normally not a problem. However the attitude of the satellite and its antenna must be carefully maintained to ensure the antenna is aligned in the correct direction. The antennas on board the satellite are typically limited in size to around 2 - 3 metres by the space that is available on the satellite structure.

For satellites in low earth orbits, considerably less directive antennas are normally used. Signals are likely to be received and transmitted over a much wider angle, and these will change as the satellites move. Accordingly these satellites seldom use parabolic reflector antennas.

Ground antennas

Ground antennas used for receing satellite signals and transmitting to the satellites vary considerably according to their application. Again parabolic reflectors are the most widely used, but Yagi antennas may be used on occasions.

The size of the antennas may vary considerably. The parabolic reflectors used for satellite television reception are very small. However those used for professional applications are much larger and may range up to several tens of metres in size.

The satellite antennas are carefully chosen by the system designer to match the particular requirements. It is possible to calculate the exact specification for the antenna, knowing the path loss, signal to noise ratio, transmitter power levels, receiver sensitivities, etc. A small 70 centimetre antenna may be sufficient for direct reception of satellite TV programmes but would not be suitable for transmitting programmes up to the satellite where a much higher signal level is required to ensure the best possible picture is radiated back to Earth.

Satellite television antennas

It has already been mentioned that satellite television antennas use parabolic reflector or "dish" antennas. They are also incorporate what is termed an LNB. This is a Low Noise Block converter. The satellite transmits signals at frequencies between 12.2 and 12.7 GHz. Signals at these frequencies would be very quickly attenuated by any coaxial feeder that was used. As feeder lengths may run into several metres or more in many installations, this would mean that the signals that reached the television would be very weak. To overcome this problem the LNB is installed at the feed point of the antenna. Its job is two fold. It amplifies the signal, but more importantly it converts it down to a frequency (usually 950 to 1450MHz) where the loss introduced by the coaxial feeder is considerably less. The amplification provided by the LNB also enables the loss introduced by the cable to be less critical. By performing these two functions it means that domestic coaxial cable can be used satisfactorily, while maintaining sufficiently high signal levels at the receiver.

Smart Antennas Tutorial

- an overview, summary, or tutorial about the basics of smart antennas or the adaptive antenna array and smart antenna technology used with SDR, cognitive radio, MIMO, and other new technologies.

Smart antennas and smart antenna technology using an adaptive antenna array are being introduced increasingly with the development of other technologies including the software defined radio, cognitive radio, MIMO and many others.

Smart antenna technology or adaptive antenna array technology enables the performance of the antenna to be altered to provide the performance that may be required to undertake performance under specific or changing conditions.

The smart antennas include signal processing capability that can perform tasks such as analysis of the direction of arrival of a signal and then the smart antenna can adapt the antenna itself using beam-forming techniques to achieve better reception, or transmission. In addition to this, the overall antenna will use some form of adaptive antenna array scheme to enable the antenna to perform is beam formation and signal direction detection.

Smart antenna functions

While the main purposes of standard antennas are to effectively transmit and receive radio signals, there are two additional functions that smart antennas or adaptive antennas need to fulfil:

- ***Direction of arrival estimation:*** In order for the smart antenna to be able provide the required functionality and optimisation of the transmission and reception, they need to be able to detect the direction of arrival of the required incoming signal. The information received by the antenna array is passed to the signal processor within the antenna and this provides the required analysis.
- ***Beam steering:*** With the direction of arrival of the required and any interfering signals analysed, the control circuitry within the antenna is able to optimise the directional beam pattern of the adaptive antenna array to provide the required performance.

Types of smart antenna

With considerable levels of functionality being required within smart antennas, two main approaches or types of smart antenna technology have been developed:

- ***Switched beam smart antennas:*** The switched beam smart or adaptive antennas are designed so that they have several fixed beam patterns. The control elements within the antenna can then select the most appropriate one for the conditions that have been detected. Although this approach does not provide complete flexibility it simplifies the design and provides sufficient level of adaptivity for many applications.
- ***Adaptive array smart antennas:*** Adaptive antenna arrays allow the beam to be continually steered to any direction to allow for the maximum signal to be received and / or the nulling of any interference.

Both types of antenna are able to provide the directivity, although decisions need to be made against cost, complexity and the performance requirements regarding which type should be used.

Summary

With many applications including MIMO, Software Defined Radio - SDR, and Cognitive Radio - CR requiring antenna systems to be more adaptive and provide greater levels of adaptivity, Smart antenna technology or adaptive antenna technology will become more widely used.

Five eighths wavelength vertical antenna

- overview, or summary about the basics of a five eigths wavelength vertical antenna, a form of RF antenna widely used for mobile radio communications applications.

Vertical antennas find widespread use in applications where an "all round" radiation pattern is required. In these applications it is necessary to keep the maximum amount of radiation parallel to the earth. It is in applications such as these that the five eighths wavelength vertical antenna has become widely used. One particular applications where they are widely used is for mobile radio communications. They are partcularly suited to mobile radio communications because there is not need for the antenna to be reorientated as the mobile station moves, and in addition to this the antenna provides gain over a quarter wave vertical.

Development

The most straightforward vertical antenna is the quarter wavelength version. However it is found that by extending the length of the vertical element, the amount of power radiated at a low angle is increased. If a half wave dipole is extended in length the radiation at right angles to the antenna starts to increase before finally splitting into several lobes. The maximum level of radiation at right angles to the antenna is achieved when the dipole is about 1.2 times the wavelength.

RF antenna gain

When used as a vertical radiator against a ground plane this translates to a length of 5/8 wavelength. It is found that a five eighths vertical has a gain of close to 4 dBd. To achieve this gain the antenna must be constructed of the right materials so that losses are reduced to the absolute minimum and the overall performance is maintained, otherwise much of the advantage of using the additional length will be lost.

Matching to the RF antenna

For most applications, it is necessary to ensure that the antenna provides a good match to 50 ohm coaxial cable. It is found that a 3/4 wavelength vertical element provides a good match, and therefore the solution to the 5/8 wavelength antenna is to make it appear as a 5/8 radiator but have the electrical length of a 3/4 element. This is achieved by placing a small loading coil at the base of the antenna to increase its electrical length.,/p>

Mechanical considerations

Five eighths wavelength vertical antennas are often used on automobiles. Accordingly one of the main constraints is to ensure that the coil at the base of the antenna is be kept rigid and does not bend as the antenna flexes with the movement of the car. If there is too much flexing then the match to the feeder will change and the operation will be impaired.

Radio Signal Path Loss

- summary, tutorial or overview about the essentials of radio signal path loss, its causes and prediction, and its use in radio coverage and wireless survey tools.

This radio path loss and link budget tutorial is split into several pages each of which address different aspects of radio path loss and link budget:

[1] Understanding signal path loss [2] Free space path loss [3] Link budget

Radio signal path loss is a particularly important element in the design of any radio communications system or wireless system. The radio signal path loss will determine many elements of the radio communications system in particular the transmitter power, and the antennas, especially their gain, height and general location. The radio path loss will also affect other elements such as the required receiver sensitivity, the form of transmission used and several other factors.

As a result, it is necessary to understand the reasons for radio path loss, and to be able to determine the levels of the signal loss for a give radio path.

The signal path loss can often be determined mathematically and these calculations are often undertaken when preparing coverage or system design activities. These depend on a knowledge of the signal propagation properties.

Accordingly, path loss calculations are used in many radio and wireless survey tools for determining signal strength at various locations. These wireless survey tools are being increasingly used to help determine what radio signal strengths will be, before installing the equipment. For cellular operators radio coverage surveys are important because the investment in a macrocell base station is high. Also, wireless survey tools provide a very valuable service for applications such as installing wireless LAN systems in large offices and other centres because they enable problems to be solved before installation, enabling costs to be considerably reduced. Accordingly there is an increasing importance being placed onto wireless survey tools and software.

Signal path loss basics

The signal path loss is essentially the reduction in power density of an electromagnetic wave or signal as it propagates through the environment in which it is travelling.

There are many reasons for the radio path loss that may occur:

- ***Free space loss:*** The free space loss occurs as the signal travels through space without any other effects attenuating the signal it will still diminish as it spreads out. This can be thought of as the radio communications signal spreading out as an ever increasing sphere. As the signal has to cover a wider area, conservation of energy tells us that the energy in any given area will reduce as the area covered becomes larger.
- ***Absorption losses:*** Absorption losses occur if the radio signal passes into a medium which is not totally transparent to radio signals. This can be likened to a light signal passing through transparent glass.
- ***Diffraction:*** Diffraction losses occur when an object appears in the path. The signal can diffract around the object, but losses occur. The loss is higher the more rounded the object. Radio signals tend to diffract better around sharp edges.
- ***Multipath:*** In a real terrestrial environment, signals will be reflected and they will reach the receiver via a number of different paths. These signals may add or subtract from each other depending upon the relative phases of the signals. If the receiver is moved the scenario will change and the overall received signal will be found vary with position. Mobile receivers (e.g. cellular telecommunications phones) will be subject to this effect which is known as Rayleigh fading.
- ***Terrain:*** The terrain over which signals travel will have a significant effect on the signal. Obviously hills which obstruct the path will considerably attenuate the signal, often making reception impossible. Additionally at low frequencies the composition of the earth will have a marked effect. For example on the Long Wave band, it is found that signals travel best over more conductive terrain, e.g. sea paths or over areas that are marshy or damp. Dry sandy terrain gives higher levels of attenuation.
- ***Buildings and vegetation:*** For mobile applications, buildings and other obstructions including vegetation have a marked effect. Not only will buildings reflect radio signals, they will also absorb them. Cellular communications are often significantly impaired within buildings. Trees and foliage can attenuate radio signals, particularly when wet.
- ***Atmosphere:*** The atmosphere can affect radio signal paths. At lower frequencies, especially below 30 - 50MHz, the ionosphere has a significant effect, reflecting (or more correctly refracting) them back to Earth. At frequencies above 50 MHz and more the troposphere has a major effect, refracting the signals back to earth as a result of changing refractive index. For UHF broadcast this can extend coverage to approximately a third beyond the horizon.

These reasons represent some of the major elements causing signal path loss for any radio system.

Predicting path loss

One of the key reasons for understanding the various elements affecting radio signal path loss is to be able to predict the loss for a given path, or to predict the coverage that may be achieved for a particular base station, broadcast station, etc.

Although prediction or assessment can be fairly accurate for the free space scenarios, for real life terrestrial applications it is not easy as there are many factors to take into consideration, and it is not always possible to gain accurate assessments of the effects they will have.

Despite this there are wireless survey tools and radio coverage prediction software programmes that are available to predict radio path loss and estimate coverage. A variety of methods are used to undertake this.

Free space path loss varies in strength as an inverse square law, i.e. 1/(range)^2, or 20 dB per decade increase in range. This calculation is very simple to implement, but real life terrestrial calculations of signal path loss are far more involved.

Most path loss predictions are made using techniques outlined below:

- ***Statistical methods:*** Statistical methods of predicting signal path loss rely on measured and averaged losses for typical types of radio links. These figures are entered into the prediction model which is able to calculate the figures based around the data. A variety of models can be used dependent upon the application. This type of approach is normally used for planning cellular networks, estimating the coverage of PMR (Private Mobile Radio) links and for broadcast coverage planning.
- ***Deterministic approach:*** This approach to radio signal path loss and coverage prediction utilises the basic physical laws as the basis for the calculations. These methods need to take into consideration all the elements within a given area and although they tend to give more accurate results, they require much additional data and computational power. In view of their complexity, they tend to be used for short range links where the amount of required data falls within acceptable limits.

These wireless survey tools and radio coverage software packages are growing in their capabilities. However it is still necessary to have a good understanding of radio propagation so that the correct figures can be entered and the results interpreted satisfactorily.

Summary

Radio signal path loss can be caused by many factors. Only in the free space scenario is the calculation straightforward. In a terrestrial environment there are many factors that affect the actual RF path loss. When planning any radio or wireless system, it is necessary to have a broad understanding the elements that give rise to the path loss, and in this way design the system accordingly.

Free Space Path Loss and Formula or Equation

- summary, tutorial or overview about the free space path loss, FSPL, and how to calculate the loss using the RF path loss formula or equation.

This radio path loss and link budget tutorial is split into several pages each of which address different aspects of radio path loss and link budget:

[1] Understanding signal path loss [2] Free space path loss [3] Link budget

The free space path loss is used in many areas for predicting radio signal strengths that may be expected in a radio system. Although it does not hold for most terrestrial situations as there are several situations in which it can be used and it is also useful as the basis for understanding many real life radio propagation situations.

Despite this, the free space path loss is an essential basic parameter for many RF calculations. It can often be used as a first approximation for many short range calculations. Alternatively it can be used as a first approximation for a number of areas where there are few obstructions. As such it is a valuable tool for many people dealing with radio communications systems.

In addition to this, these calculations can be used in wireless survey tools. With the growing requirements to be able to analyse wireless or radio coverage, wireless survey tools are being sued increasingly to enable coverage to be predicted at the early stages of design. Accordingly these wireless survey tools are being used increasingly in the development and installation of radio and wireless systems.

Free space path loss basics

The free space path loss, also known as FSPL is the loss in signal strength that occurs when an electromagnetic wave travels over a line of sight path in free space. In these circumstances there are no obstacles that might cause the signal to be reflected refracted, or that might cause additional attenuation.

The free space path loss calculations only look at the loss of the path itself and do not contain any factors relating to the transmitter power, antenna gains or the receiver sensitivity levels. These factors are normally address when calculating a link budget and these will also be used within radio and wireless survey tools and software.

To understand the reasons for the free space path loss, it is possible to imagine a signal spreading out from a transmitter. It will move away from the source spreading out in the form of a sphere. As it does so, the surface area of the sphere increases. As this will follow the law of the conservation of energy, as the surface area of the sphere increases, so the intensity of the signal must decrease.

As a result of this it is found that the signal decreases in a way that is inversely proportional to the square of the distance from the source of the radio signal.

$$\textbf{Signal} = \frac{\textbf{1}}{\textbf{distance}^2}$$

Free space path loss formula

The free space path loss formula or free space path loss equation is quite simple to use. Not only is the path loss proportional to the square of the distance between the transmitter and receiver, but the signal level is also proportional to the square of the frequency in use for other reasons explained in a section below.

$$\textbf{FSPL} = (4 \pi d / \lambda)^2$$

$$= (4 \pi d f / c)^2$$

Where:
FSPL is the Free space path loss
d is the distance of the receiver from the transmitter (metres)
λ is the signal wavelength (metres)
f is the signal frequency (Hertz)
c is the speed of light in a vacuum (metres per second)

The speed of light is 2.99792458×10^8 metres per second, although for most practical purposes, this is taken to be 3×10^8 metres per second.

The free space path loss formula is applicable to situations where only the electromagnetic wave is present, i.e. for far field situations. It does not hold true for near field situations.

Free space loss formula frequency dependency

Although the free space loss equation given above seems to indicate that the loss is frequency dependent. The attenuation provided by the distance travelled in space is not dependent upon the frequency. This is constant.

The reason for the frequency dependence is that the equation contains two effects:

1. The first results from the spreading out of the energy as the sphere over which the energy is spread increases in area. This is described by the inverse square law.

2. The second effect results from the antenna aperture change. This affects the way in which any antenna can pick up signals and this term is frequency dependent.

As one constituent of the path loss equation is frequency dependent, this means that there is a frequency dependency within the complete equation.

Decibel version of free space path loss equation

Most RF comparisons and measurements are performed in decibels. This gives an easy and consistent method to compare the signal levels present at various points. Accordingly it is very convenient to express the free space path loss formula, FSPL, in terms of decibels. It is easy to take the basic free space path loss equation and manipulate into a form that can be expressed in a logarithmic format.

$$\mathbf{FSPL\ (dB) = 20\ log_{10}\ (d) + 20\ log_{10}\ (f) + 32.44}$$

Where:
d is the distance of the receiver from the transmitter (km)
f is the signal frequency (MHz)

Affect of antenna gain on path loss equation

The equation above does not include any component for antenna gains. It is assumed that the antenna gain is unity for both the transmitter. In reality, though, all antennas will have a certain amount of gain and this will affect the overall affect. Any antenna gain will reduce the "loss" when compared to a unity gain system. The figures for antenna gain are relative to an isotropic source, i.e. an antenna that radiates equally in all directions.

$$\mathbf{FSPL\ (dB) = 20\ log_{10}\ (d) + 20\ log_{10}\ (f) + 32.44\ -Gtx\ -\ Grx}$$

Where:
Gtx is the gain of the transmitter antenna relative to an isotropic source (dBi)
Grx is the gain of the receiver antenna relative to an isotropic source (dBi)

Summary

The free space path loss equation or formula given above, is an essential tool that is required when making calculations for radio and wireless systems either manually or within applications such as wireless survey tools, etc. By using the free space path loss equation, it is possible to determine the signal strengths that may be expected in many scenarios. While the free space path loss formula is not fully applicable where there are other interactions, e.g. reflection, refraction, etc as are present in most real life applications, the equation can nevertheless be used to give an indication of what may be expected. It is obviously fully applicable to satellite systems where the paths conform closely to the totally free space scenarios.

Link Budget

- summary, tutorial or overview about the essentials of link budget and how to calculate the link budget for satellite and other radio systems.

This radio path loss and link budget tutorial is split into several pages each of which address different aspects of radio path loss and link budget:

[1] Understanding signal path loss [2] Free space path loss [3] Link budget

When designing a complete, i.e. end to end radio communications system, it is necessary to calculate what is termed the link budget. The link budget enables factors such as the required antennas gain levels, radio transmitter power levels, and receiver sensitivity figures to be determined. By assessing the link budget, it is possible to design the system so that it meets its requirements and performs correctly without being over designed at extra cost.

Link budgets are often used for satellite systems. In these situations it is crucial that the required signal levels are maintained to ensure that the received signal levels are sufficiently high above the noise level to ensure that signal to noise levels or bit error rates are within the required limits. However larger antennas, high transmitter power levels that required add considerably to the cost, so it is necessary to balance these to minimise the cost of the system while still maintaining performance.

In addition to satellite systems, link budgets are also used in many other radio communications systems. For example, link budget calculations are used for calculating the power levels required for cellular communications systems, and for investigating the base station coverage.

Link budget style calculations are also used within wireless survey tools. These wireless survey tools will not only look at the way radio signals propagate, but also the power levels, antennas and receiver sensitivity levels required to provide the required link quality.

What is link budget?

As the name implies, a link budget is an accounting of all the gains and losses in a transmission system. The link budget looks at the elements that will determine the signal strength arriving at the receiver. The link budget may include the following items:

- Transmitter power.
- Antenna gains (receiver and transmitter).
- Antenna feeder losses (receiver and transmitter).
- Path losses.
- Receiver sensitivity (although this is not part of the actual link budget, it is necessary to know this to enable any pass fail criteria to be applied.

Where the losses may vary with time, e.g. fading, and allowance must be made within the link budget for this - often the worst case may be taken, or alternatively an acceptance of periods of increased bit error rate (for digital signals) or degraded signal to noise ratio for analogue systems.

In essence the link budget will take the form of the equation below:

Received power (dBm) = Transmitted power (dBm) + gains (db) - losses (dB)

The basic calculation to determine the link budget is quite straightforward. It is mainly a matter of accounting for all the different losses and gains between the transmitter and the receiver.

Link budget equation

In order to devise a link budget equation, it is necessary to investigate all the areas where gains and losses may occur between the transmitter and the receiver. Although guidelines and suggestions can be made regarding the possible areas for losses and gains, each link has to be analysed on its own merits..

A typical link budget equation for a radio communications system may look like the following:

$$P_{RX} = P_{TX} + P_{TX} + G_{TX} + G_{RX} - L_{TX} - L_{FS} - L_P - L_{RX}$$

Where:

P_{RX} = received power (dBm)
P_{TX} = transmitter output power (dBm)

G_{TX} = transmitter antenna gain (dBi)
G_{RX} = receiver antenna gain (dBi)
L_{TX} = transmit feeder and associated losses (feeder, connectors, etc.) (dB)
L_{FS} = free space loss or path loss (dB)
L_{P} = miscellaneous signal propagation losses (these include fading margin, polarization mismatch, losses associated with medium through which signal is travelling, other losses...) (dB)
L_{RX} = receiver feeder and associated losses (feeder, connectors, etc.) (d)B

NB for the sake of showing losses in the link budget equation is "minus" actual loss figures, e.g. LTX or LFS, etc should be taken as the modulus of the loss.

Antenna gain and link budget

The basic link budget equation where no levels of antenna gain are included assumes that the power spreads out equally in all directions from the source. In other words the antenna is an isotropic source, radiating equally in all directions.

This assumption is good for theoretical calculations, but in reality all antennas radiate more in some directions than others. In addition to this it is often necessary to use antennas with gain to enable interference from other directions to be reduced at the receiver, and at the transmitter to focus the available transmitter power in the required direction.

In view of this it is necessary to accommodate these gains into the link budget equation as they have been in the equation above because they will affect the signal levels - increasing them by levels of the antenna gain, assuming the gain is in the direction of the required link.

When quoting gain levels for antennas it is necessary to ensure they are gains when compared to an isotropic source, i.e. the basic type of antenna assumed in the equation when no gain levels are incorporated. The gain figures relative to an isotropic source are quoted as dBi, i.e. dB relative to an isotropic source. Often gain levels given for an antenna may be the gain relative to a dipole where the figures may be quoted as dBd, i.e. dB relative to a dipole. However a dipole has gain relative to an isotropic source, so the dipole gain of 2.1 dBi needs to be accommodated if figures relative to a dipole are quoted for an antenna gain..

Effect of multipath propagation

For true free space propagation such as that encountered for satellites there will be no noticeable reflections and there will only be one major path. However for terrestrial systems, the signal may reach the receiver via a number of different paths as a result of reflections, etc that will occur as a result of the objects around the path. Buildings, trees, objects around the office and home can all cause reflections that will result in the signal variations.

The multipath propagation will cause variations of the signal strength when compared to that calculated from the free space path loss. If the signals arrive in phase with the direct signal, then the reflected signals will tend to reinforce the direct signal. If they are out of phase, then they will tend to cancel the signal. If either the transmitter or receiver moves, then the signal strength will be seen to vary as the relative strengths and phases of the different signals change.

In order to allow for this in a link budget, a link margin is added into the equation to allow for this.

Summary

Link budget calculations are an essential step in the deign of a radio communications system. The link budget calculation enables the losses and gains to be seen, and devising a link budget enables the apportionment of losses, gains and power levels to be made if changes need to be made to enable the radio communications system to meet its operational requirements. Only by performing a link budget analysis is this possible.

- **Home**

Navigation:: Home >> Antennas and propagation >> this page

Link Budget

- summary, tutorial or overview about the essentials of link budget and how to calculate the link budget for satellite and other radio systems.

This radio path loss and link budget tutorial is split into several pages each of which address different aspects of radio path loss and link budget:

[1] Understanding signal path loss [2] Free space path loss [3] Link budget

When designing a complete, i.e. end to end radio communications system, it is necessary to calculate what is termed the link budget. The link budget enables factors such as the required antennas gain levels, radio transmitter power levels, and receiver sensitivity figures to be determined. By assessing the link budget, it is possible to design the system so that it meets its requirements and performs correctly without being over designed at extra cost.

Link budgets are often used for satellite systems. In these situations it is crucial that the required signal levels are maintained to ensure that the received signal levels are sufficiently high above the noise level to ensure that signal to noise levels or bit error rates are within the required limits. However larger antennas, high transmitter power levels that required add considerably to the cost, so it is necessary to balance these to minimise the cost of the system while still maintaining performance.

In addition to satellite systems, link budgets are also used in many other radio communications systems. For example, link budget calculations are used for calculating the power levels required for cellular communications systems, and for investigating the base station coverage.

Link budget style calculations are also used within wireless survey tools. These wireless survey tools will not only look at the way radio signals propagate, but also the power levels, antennas and receiver sensitivity levels required to provide the required link quality.

What is link budget?

As the name implies, a link budget is an accounting of all the gains and losses in a transmission system. The link budget looks at the elements that will determine the signal strength arriving at the receiver. The link budget may include the following items:

- Transmitter power.
- Antenna gains (receiver and transmitter).
- Antenna feeder losses (receiver and transmitter).
- Path losses.
- Receiver sensitivity (although this is not part of the actual link budget, it is necessary to know this to enable any pass fail criteria to be applied.

Where the losses may vary with time, e.g. fading, and allowance must be made within the link budget for this - often the worst case may be taken, or alternatively an acceptance of periods of increased bit error rate (for digital signals) or degraded signal to noise ratio for analogue systems.

In essence the link budget will take the form of the equation below:

Received power (dBm) = Transmitted power (dBm) + gains (db) - losses (dB)

The basic calculation to determine the link budget is quite straightforward. It is mainly a matter of accounting for all the different losses and gains between the transmitter and the receiver.

Link budget equation

In order to devise a link budget equation, it is necessary to investigate all the areas where gains and losses may occur between the transmitter and the receiver. Although guidelines and suggestions can be made regarding the possible areas for losses and gains, each link has to be analysed on its own merits..

A typical link budget equation for a radio communications system may look like the following:

$$\mathbf{P_{RX} = P_{TX} + P_{TX} + G_{TX} + G_{RX} - L_{TX} - L_{FS} - L_{P} - L_{RX}}$$

Where:
P_{RX} = received power (dBm)
P_{TX} = transmitter output power (dBm)
G_{TX} = transmitter antenna gain (dBi)
G_{RX} = receiver antenna gain (dBi)
L_{TX} = transmit feeder and associated losses (feeder, connectors, etc.) (dB)
L_{FS} = free space loss or path loss (dB)
L_{P} = miscellaneous signal propagation losses (these include fading margin, polarization mismatch, losses associated with medium through which signal is travelling, other losses...) (dB)
L_{RX} = receiver feeder and associated losses (feeder, connectors, etc.) (d)B

NB for the sake of showing losses in the link budget equation is "minus" actual loss figures, e.g. LTX or LFS, etc should be taken as the modulus of the loss.

Antenna gain and link budget

The basic link budget equation where no levels of antenna gain are included assumes that the power spreads out equally in all directions from the source. In other words the antenna is an isotropic source, radiating equally in all directions.

This assumption is good for theoretical calculations, but in reality all antennas radiate more in some directions than others. In addition to this it is often necessary to use antennas with gain to enable interference from other directions to be reduced at the receiver, and at the transmitter to focus the available transmitter power in the required direction.

In view of this it is necessary to accommodate these gains into the link budget equation as they have been in the equation above because they will affect the signal levels - increasing them by levels of the antenna gain, assuming the gain is in the direction of the required link.

When quoting gain levels for antennas it is necessary to ensure they are gains when compared to an isotropic source, i.e. the basic type of antenna assumed in the equation when no gain levels are incorporated. The gain figures relative to an isotropic source are quoted as dBi, i.e. dB relative to an isotropic source. Often gain levels given for an antenna may be the gain relative to a dipole where the figures may be quoted as dBd, i.e. dB relative to a dipole. However a dipole has gain relative to an isotropic source, so the dipole gain of 2.1 dBi needs to be accommodated if figures relative to a dipole are quoted for an antenna gain..

Effect of multipath propagation

For true free space propagation such as that encountered for satellites there will be no noticeable reflections and there will only be one major path. However for terrestrial systems, the signal may reach the receiver via a number of different paths as a result of reflections, etc that will occur as a result of the objects around the path. Buildings, trees, objects around the office and home can all cause reflections that will result in the signal variations.

The multipath propagation will cause variations of the signal strength when compared to that calculated from the free space path loss. If the signals arrive in phase with the direct signal, then the reflected signals will tend to reinforce the direct signal. If they are out of phase, then they will tend to cancel the signal. If either the transmitter or receiver moves, then the signal strength will be seen to vary as the relative strengths and phases of the different signals change.

In order to allow for this in a link budget, a link margin is added into the equation to allow for this.

Summary

Link budget calculations are an essential step in the deign of a radio communications system. The link budget calculation enables the losses and gains to be seen, and devising a link budget enables the apportionment of losses, gains and power levels to be made if changes need to be made to enable the radio communications system to meet its operational requirements. Only by performing a link budget analysis is this possible.

- **Home**
-
- Antennas & propagation
- Broadcast technology
- Cellular telecoms
- Circuits
- Design principles
- Electronic components
- Electronics manufacture
- Formulae & data
- Radio receivers
- RF topics
- Satellite technology
- Technology history

Multipath Fading

- summary, tutorial or overview about the essentials multipath fading, a phenomenon that is present on many radio signals, cellular, HF and VHF.

This multipath propagation and fading tutorial is split into several pages each of which address different aspects of multipath propagation and fading:

[1] Multipath propagation basics [2] Multipath fading [3] Rayleigh fading

Multipath fading affects most forms of radio communications links in one form or another. Multipath fading can be detected on many signals across the frequency spectrum from the HF bands right up to microwaves and beyond. It is experienced not only by short wave radio communications where signals fade in and out over a period of time, but it is also experienced by many other forms of radio communications systems including cellular telecommunications and many other users of the VHF and UHF spectrum.

Multipath fading occurs in any environment where there is multipath propagation and there is some movement of elements within the radio communications system. This may include the radio transmitter or receiver position, or in the elements that give rise to the reflections. The multipath fading can often be relatively deep, i.e. the signals fade completely away, whereas at other times the fading may not cause the signal to fall below a useable strength.

Multipath fading may also cause distortion to the radio signal. As the various paths that can be taken by the signals vary in length, the signal transmitted at a particular instance will arrive at the receiver over a spread of times. This can cause problems with phase distortion and intersymbol interference when data transmissions are made. As a result, it may be necessary to incorporate features within the radio communications system that enables the effects of these problems to be minimised.

Multipath fading basics

Multipath fading is a feature that needs to be taken into account when designing or developing a radio communications system. In any terrestrial radio communications system, the signal will reach the receiver not only via the direct path, but also as a result of reflections from objects such as buildings, hills, ground, water, etc that are adjacent to the main path.

The overall signal at the radio receiver is a summation of the variety of signals being received. As they all have different path lengths, the signals will add and subtract from the total dependent upon their relative phases.

At times there will be changes in the relative path lengths. This could result from either the radio transmitter or receiver moving, or any of the objects that provides a reflective surface moving. This will result in the phases of the signals arriving at the receiver changing, and in turn this will result in the signal strength varying as a result of the different way in which the signals will sum together. It is this that causes the fading that is present on many signals.

Selective and flat fading

Multipath fading can affect radio communications channels in two main ways. This can given the way in which the effects of the multipath fading are mitigated.

1. ***Flat fading:*** This form of multipath fading affects all the frequencies across a given channel either equally or almost equally. When flat multipath fading is experienced, the signal will just change in amplitude, rising and falling over a period of time, or with movement from one position to another.

2. ***Selective fading:*** Selective fading occurs when the multipath fading affects different frequencies across the channel to different degrees. It will mean that the phases and amplitudes of the signal will vary across the channel. Sometimes relatively deep nulls may be experienced, and this can give rise to some reception problems. Simply maintaining the overall amplitude of the received signal will not overcome the effects of selective fading, and some form of equalization may be needed. Some digital signal formats, e.g. OFDM are able to spread the data over a wide channel so that only a portion of the data is lost by any nulls. This can be reconstituted using forward error correction techniques and in this way it can mitigate the effects of selective multipath fading.

   Selective multipath fading occurs because even though the path length will be change by the same physical length (e.g. the same number of metres, yards, miles, etc) this represents a different proportion of a wavelength. Accordingly the phase will change across the bandwidth used.

Selective fading can occur over many frequencies. It can often be noticed when medium wave broadcast stations are received in the evening via ground wave and skywave. The phases of the signals received via the two means of propagation change with time and this causes the overall received signal to change. As the multipath fading is very dependent on path length, it is found that it affects the frequencies over even the bandwidth of an AM broadcast signal to be affected differently and distortion results.

Selective multipath fading is also experienced at higher frequencies, and with high data rate signals becoming commonplace wider bandwidths are needed. As a result nulls and peaks may occur across the bandwidth of a single signal.

Cellular multipath fading

Cellular telecommunications is subject to multipath fading. There are a variety of reasons for this. The first is that the mobile station or user is likely to be moving, and as a result the path lengths of all the signals being received are changing. The second is that many objects around may also be moving. Automobiles and even people will cause reflections that will have a significant effect on the received signal. Accordingly multipath fading has a major bearing on cellular telecommunications.

Often the multipath fading that affects cellular phones is known as fast fading because it occurs over a relatively short distance. Slow fading occurs as a cell phone moves behind an obstruction and the signal slowly fades out.

The fast signal variations caused by multipath fading can be detected even over a short distance. Assume a frequency of 2 GHz (e.g. a typical approximate frequency value for many 3G phones). The wavelength can be calculated as:

$$\begin{aligned} \boldsymbol{\lambda} &= \mathbf{c \;/\; f} \\ &= \mathbf{3 \times 10^{8} \;/\; 2 \times 10^{9}} \\ &= \mathbf{0.15\ metres} \end{aligned}$$

Where:
c = speed of light in metres per second
f = frequency in Hertz

To move from a signal being in phase to a signal being out of phase is equivalent to increasing the path length by half a wavelength or 0.075m, or 7.5 cms. This example looks at a very simplified example. In reality the situation is far more complicated with signals being received via many paths. However it does give an indication of the distances involved to change from an in-phase to an out of phase situation.

Ionospheric multipath fading

Short wave radio communications is renowned for its fading. Signals that are reflected via the ionosphere, vary considerably in signal strength. These variations in strength are primarily caused by multipath fading.

When signals are propagated via the ionosphere it is possible for the energy to be propagated from the transmitter to the receiver via very many different paths. Simple diagrams show a single ray or path that the signal takes. In reality the profile of the electron density of the ionosphere (it is the electron density profile that causes the signals to be refracted) is not smooth and as a result any signals entering the ionosphere will be scattered and will take a variety of paths to reach a particular receiver. With changes in the ionosphere causing the path lengths to change, this will result in the phases changing and the overall summation at the receiver changing. [See the pages on ionospheric propagation within the Radio Wave Propagation section of this website for further details of this form of propagation].

The changes in the ionosphere arise from a number of factors. One is that the levels of ionisation vary, although these changes normally occur relatively slowly, but nevertheless have an effect. In addition to this there are winds or air movements in the ionosphere. As the levels of ionisation are not constant, any air movement will cause changes in the profile of the electron density in the ionosphere. In turn this will affect the path lengths.

Tropospheric multipath fading

Many signals using frequencies at VHF and above are affected by the troposphere. The signal is refracted as a result of the changes in refractive index occurring, especially within the first kilometres above the ground. This can cause signals to travel beyond the line of sight. In fact for broadcast applications a figure of 4/3 of the visual line of sight is used for the radio horizon. However under some circumstances relatively abrupt changes in refractive index occurring as a result of weather conditions can cause the distances over which signals travel to be increased. Signals may then be "ducted" by the ionosphere over distances up to a few hundred kilometres. [See the pages on tropospheric propagation within the Radio Wave Propagation section of this website for further details of this form of propagation].

When signals are ducted in this way, they will be subject to multipath fading. Here, heat rising from the Earth's surface will ensure that the path is always changing and signals will vary in strength. Typically these changes may be relatively slow with signals falling and rising in strength over a period of a number of minutes.

Summary

Multipath fading is a feature of many radio communications links. Multipath fading occurs as a result of the many signal paths that are in existence on all terrestrial radio communications links whether they are used for applications such as cellular telecommunications, mobile radio, or for HF or VHF radio communications. As such it is necessary to account for multipath fading in the design of many radio communications systems.

Rayleigh Fading

- summary, tutorial or overview about the essentials of Rayleigh fading and how it applies to radio signal paths where multiple reflections and paths are available.

This multipath propagation and fading tutorial is split into several pages each of which address different aspects of multipath propagation and fading:

[1] Multipath propagation basics [2] Multipath fading [3] Rayleigh fading

Rayleigh fading is the name given to the form of fading that is often experienced in an environment where there is a large number of reflections present. The Rayleigh fading model uses a statistical approach to analyse the propagation, and can be used in a number of environments.

The Rayleigh fading model is normally viewed as a suitable approach to take when analysing and prediction radio wave propagation performance for areas such as cellular communications in a well built up urban environment where there are many reflections from buildings, etc.. HF ionospheric radio wave propagation where reflections (or more exactly refractions) occur at many points within the ionosphere is also another area where Rayleigh fading model applies well. It is also appropriate to use the Rayleigh fading model for tropospheric radio propagation because, again there are many reflection points and the signal may follow a variety of different paths.

The Rayleigh propagation model is most applicable to instances where there are many different signal paths, none of which is dominant. In this way all the signal paths will vary and can have an impact on the overall signal at the receiver.

Rayleigh fading basics

The Rayleigh fading model is particularly useful in scenarios where the signal may be considered to be scattered between the transmitter and receiver. In this form of scenario there is no single signal path that dominates and a statistical approach is required to the analysis of the overall nature of the radio communications channel.

Rayleigh fading is a model that can be used to describe the form of fading that occurs when multipath propagation exists. In any terrestrial environment a radio signal will travel via a number of different paths from the transmitter to the receiver. The most obvious path is the direct, or line of sight path.

However there will be very many objects around the direct path. These objects may serve to reflect, refract, etc the signal. As a result of this, there are many other paths by which the signal may reach the receiver.

When the signals reach the receiver, the overall signal is a combination of all the signals that have reached the receiver via the multitude of different paths that are available. These signals will all sum together, the phase of the signal being important. Dependent upon the way in which these signals sum together, the signal will vary in strength. If they were all in phase with each other they would all add together. However this is not normally the case, as some will be in phase and others out of phase, depending upon the various path lengths, and therefore some will tend to add to the overall signal, whereas others will subtract.

As there is often movement of the transmitter or the receiver this can cause the path lengths to change and accordingly the signal level will vary. Additionally if any of the objects being used for reflection or refraction of any part of the signal moves, then this too will cause variation. This occurs because some of the path lengths will change and in turn this will mean their relative phases will change, giving rise to a change in the summation of all the received signals.

The Rayleigh fading model can be used to analyse radio signal propagation on a statistical basis. It operates best under conditions when there is no dominant signal (e.g. direct line of sight signal), and in many instances cellular telephones being used in a dense urban environment fall into this category. Other examples where no dominant path generally exists are for ionospheric propagation where the signal reaches the receiver via a huge number of individual paths. Propagation using tropospheric ducting also exhibits the same patterns. Accordingly all these examples are ideal for the use of the Rayleigh fading or propagation

Ground Wave Signal Propagation

- the ground wave used for radio communications signal propagation on the long, and medium wave bands for local radio communications

Ground wave propagation is particularly important on the LF and MF portion of the radio spectrum. Ground wave radio propagation is used to provide relatively local radio communications coverage, especially by radio broadcast stations that require to cover a particular locality.

Ground wave radio signal propagation is ideal for relatively short distance propagation on these frequencies during the daytime. Sky-wave ionospheric propagation is not possible during the day because of the attenuation of the signals on these frequencies caused by the D region in the ionosphere. In view of this, radio communications stations need to rely on the ground-wave propagation to achieve their coverage.

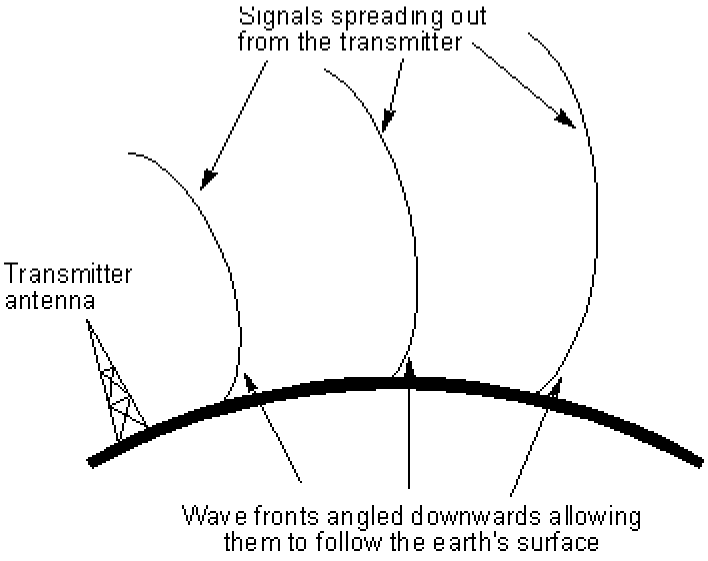
A ground wave radio signal is made up from a number of constituents. If the antennas are in the line of sight then there will be a direct wave as well as a reflected signal. As the names suggest the direct signal is one that travels directly between the two antenna and is not affected by the locality. There will also be a reflected signal as the transmission will be reflected by a number of objects including the earth's surface and any hills, or large buildings. That may be present.

In addition to this there is surface wave. This tends to follow the curvature of the Earth and enables coverage to be achieved beyond the horizon. It is the sum of all these components that is known as the ground wave.

Beyond the horizon the direct and reflected waves are blocked by the curvature of the Earth, and the signal is purely made up from the diffracted surface wave. It is for this reason that surface wave is commonly called ground wave propagation.

Surface wave

The radio signal spreads out from the transmitter along the surface of the Earth. Instead of just travelling in a straight line the radio signals tend to follow the curvature of the Earth. This is because currents are induced in the surface of the earth and this action slows down the wave-front in this region, causing the wave-front of the radio communications signal to tilt downwards towards the Earth. With the wave-front tilted in this direction it is able to curve around the Earth and be received well beyond the horizon.


Signals spreading out
from the transmitter
Transmitter
antenna
Wave fronts angled downwards allowing
them to follow the earth's surface

Ground wave radio propagation

Effect of frequency

As the wavefront of the ground wave travels along the Earth's surface it is attenuated. The degree of attenuation is dependent upon a variety of factors. Frequency of the radio signal is one of the major determining factor as losses rise with increasing frequency. As a result it makes this form of propagation impracticable above the bottom end of the HF portion of the spectrum (3 MHz). Typically a signal at 3.0 MHz will suffer an attenuation that may be in the region of 20 to 60 dB more than one at 0.5 MHz dependent upon a variety of factors in the signal path including the distance. In view of this it can be seen why even high power HF radio broadcast stations may only be audible for a few miles from the transmitting site via the ground wave.

Effect of the ground

The surface wave is also very dependent upon the nature of the ground over which the signal travels. Ground conductivity, terrain roughness and the dielectric constant all affect the signal attenuation. In addition to this the ground penetration varies, becoming greater at lower frequencies, and this means that it is not just the surface conductivity that is of interest. At the higher frequencies this is not of great importance, but at lower frequencies penetration means that ground strata down to 100 metres may have an effect.

Despite all these variables, it is found that terrain with good conductivity gives the best result. Thus soil type and the moisture content are of importance. Salty sea water is the best, and rich agricultural, or marshy land is also good. Dry sandy terrain and city centres are by far the worst. This means sea paths are optimum, although even these are subject to variations due to the roughness of the sea, resulting on path losses being slightly dependent upon the weather! It should also be noted that in view of the fact that signal penetration has an effect, the water table may have an effect dependent upon the frequency in use.

Effect of polarisation

The type of antenna has a major effect. Vertical polarisation is subject to considerably less attenuation than horizontally polarised signals. In some cases the difference can amount to several tens of decibels. It is for this reason that medium wave broadcast stations use vertical antennas, even if they have to be made physically short by adding inductive loading. Ships making use of the MF marine bands often use inverted L antennas as these are able to radiate a significant proportion of the signal that is vertically polarised.

At distances that are typically towards the edge of the ground wave coverage area, some sky-wave signal may also be present, especially at night when the D layer attenuation is reduced.

This may serve to reinforce or cancel the overall signal resulting in figures that will differ from those that may be expected.

Tropospheric propagation

- overview, tutorial or article about tropospheric propagation and the way in which radio signals at VHF and above propagate over greater distances than strictly line of sight in the troposphere

On frequencies above 30 MHz, it is found that the troposphere has an increasing effect on radio signals and radio communications systems. The radio signals are able to travel over greater distances than would be suggested by line of sight calculations. At times conditions change and radio signals may be detected over distances of 500 or even 1000 km and more. This is normally by a form of tropospheric enhancement, often called "tropo" for short. At times signals may even be trapped in an elevated duct in a form of radio signal propagation known as tropospheric ducting. This can disrupt many radio communications links (including two way radio communications links) because interference may be encountered that is not normally there. As a result when designing a radio communications link or network, this form of interference must be recognised so that steps can be taken to minimise its effects.

The way in which signals travel at frequencies of VHF and above is of great importance for those looking at radio coverage of systems such as cellular telecommunications, mobile radio communications and other wireless systems as well as other users including radio hams.

Line of sight radio communications

It might be thought that most radio communications links at VHF and above follow a line of sight path. This is not strictly true and it is found that even under normal conditions radio signals are able to travel or propagate over distances that are greater than the line of sight.

The reason for the increase in distance travelled by the radio signals is that they are refracted by small changes that exist in the Earth's atmosphere close to the ground. It is found that the refractive index of the air close to the ground is very slightly higher than that higher up. As a result the radio signals are bent towards the area of higher refractive index, which is closer to the ground. It thereby extends the range of the radio signals.

The refractive index of the atmosphere varies according to a variety of factors. Temperature, atmospheric pressure and water vapour pressure all influence the value. Even small changes in these variables can make a significant difference because radio signals can be refracted over whole of the signal path and this may extend for many kilometres.

N units

It is found that the average value for the refractive index of air at ground level is around 1.0003, but it can easily vary from 1.00027 to 1.00035. In view of the very small changes that are seen, a system has been introduced that enables the small changes to be noted more easily. Units called "N" units are often used. These N-units are obtained by subtracting 1 from the refractive index and multiply the remainder by a million. In this way more manageable numbers are obtained.

N = (mu-1) x 10^6

Where mu is the refractive index

It is found that as a very rough guide under normal conditions in a temperature zone, the refractive index of the air falls by about 0.0004 for every kilometre increase in height, i.e. 400 N units / km. This causes the radio signals to tend to follow the earth's curvature and travel beyond the geometric horizon. The actual values extend the radio horizon by about a third. This factor is often used in most radio communications coverage calculations for applications such as broadcast radio transmitters, and other two way radio communications users such as mobile radio communications, cellular telecommunications and the like.

Enhanced conditions

Under certain conditions the radio propagation conditions provided by the troposphere are such that signals travel over even greater distances. This form of "lift" in conditions is less pronounced on the lower portions of the VHF spectrum, but is more apparent on some of the higher frequencies. Under some conditions radio signals may be heard over distances of 2000 or more kilometres with distances of 3000 kilometres being possible on rare occasions. This can give rise to significant levels of interference for periods of time.

These extended distances result from much greater changes in the values of refractive index over the signal path. This enables the signal to achieve a greater degree of bending and as a result follow the curvature of the Earth over greater distances.

Under some circumstances the change in refractive index may be sufficiently high to bend the signals back to the Earth's surface at which point they are reflected upwards again by the Earth's surface. In this way the signals may travel around the curvature of the Earth, being reflected by its surface. This is one form of "tropospheric duct" that can occur.

It is also possible for tropospheric ducts to occur above the Earth's surface. These elevated tropospheric ducts occur when a mass of air with a high refractive index has a mass of air with a lower refractive index underneath and above it as a result of the movement of air that can occur under some conditions. When these conditions occur the signals may be confined within the elevated area of air with the high refractive index and they cannot escape and return to earth. As a result they may travel for several hundred miles, and receive comparatively low levels of attenuation. They may also not audible to stations underneath the duct and in this way create a skip or dead zone similar to that experienced with HF ionospheric propagation.

Mechanism behind tropospheric propagation

Tropospheric propagation effects occur comparatively close to the surface of the Earth. The radio signals are affected by the region that is below an altitude of about 2 kilometres. As these regions are those that are greatly affected by the weather, there is a strong link between weather conditions and radio propagation conditions and coverage.

Under normal conditions a there is a steady gradient of the refractive index with height, the air being closest to the Earth's surface having the highest refractive index. This is caused by several factors. Air having a higher density and that containing a higher concentration of water vapour both lead to an increase in refractive index. As the air closest to the Earth's surface is both more dense (as a result of the pressure exerted by the gases above it) and has a higher concentration of water vapour than that higher up mean that the refractive index of the air closest to the earth's surface is the highest.

Normally the temperature of the air closest to the Earth's surface is higher than that at a greater altitude. This effect tends to reduce the air density gradient (and hence the refractive index gradient) as air with a higher temperature is less dense.

However, under some circumstances, what is termed a temperature inversion occurs. This happens when the hot air close to the earth rises allowing colder denser air to come in close to the Earth. When this occurs it gives rise to a greater change in refractive index with height and this results in a more significant change in refractive index.

Temperature inversions can arise in a number of ways. One of the most dramatic occurs when an area of high pressure is present. A high pressure area means that stable weather conditions will be present, and during the summer they are associated with warm weather. The conditions mean that air close to the ground heats up and rises. As this happens colder air flows in underneath it causing the temperature inversion. Additionally it is found that the greatest improvements tend to occur as the high-pressure area is moving away and the pressure is just starting to fall.

A temperature inversion may also occur during the passage of a cold front. A cold front occurs when an area of cold air meets an area of warm air. Under these conditions the warm air rises above the cold air creating a temperature inversion. Cold fronts tend to move relatively quickly and as a result the improvement in propagation conditions tends to be short lived.

Fading

When signals are propagated over extended distances as a result of enhanced tropospheric propagation conditions, the signals are normally subject to slow deep fading. This is caused by the fact that the signals are received via a number of different paths. As the winds in the atmosphere move the air around it means that the different paths will change over a period of time. Accordingly the signals appearing at the receiver will fall in and out of phase with each other as a result of the different and changing path lengths, and as a result the strength of the overall received signal will change.

Summary

Any terrestrial signals received at VHF and above will be subject to the prevailing propagation conditions caused by the troposphere. Under normal conditions it should be expected that signals will be able to be received beyond the normal line of sight distance. However under some circumstances these distances will be considerably increased and significant levels of interference may be experienced.

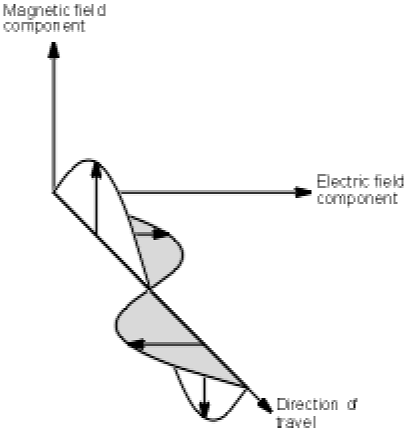
Electromagnetic waves and radio propagation

This radio path loss and link budget tutorial is split into several pages each of which address different aspects of radio path loss and link budget:

[1] Electromagnetic waves [2] Reflection, refraction, diffraction of e/m waves [3] Polarisation of e/m waves [4] Electromagnetic spectrum

Radio signals exist as a form of electromagnetic wave. These radio signals are the same form of radiation as light, ultra-violet, infra-red, etc., differing only in the wavelength or frequency of the radiation.

Electromagnetic waves have two elements. They are made from electric and magnetic components that are inseparable. The planes of the fields are at right angles to each other and to the direction in which the wave is travelling.



An electromagnetic wave

It is useful to see where the different elements of the wave emanate from to gain a more complete understanding of electromagnetic waves. The electric component of the wave results from the voltage changes that occur as the antenna element is excited by the alternating waveform. The lines of force in the electric field run along the same axis as the antenna, but spreading out as they move away from it. This electric field is measured in terms of the change of potential over a given distance, e.g. volts per metre, and this is known as the field strength. This measure is often used in measuring the intensity of an electromagnetic wave at a particular point. The other component, namely the magnetic field is at right angles to the electric field and hence it is at right angles to the plane of the antenna. It is generated as a result of the current flow in the antenna.

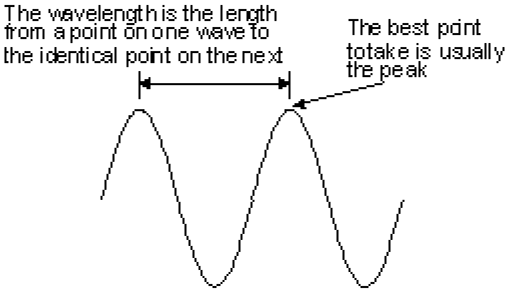
Like other forms of electromagnetic wave, radio signals can be reflected, refracted and undergo diffraction. In fact some of the first experiments with radio waves proved these facts, and they were used to establish a link between radio waves and light rays.

Wavelength, frequency and velocity

There are a number of basic properties of electromagnetic waves, or any repetitive waves for that matter that are particularly important.

One of the first that is quoted is their speed. Radio waves travel at the same speed as light. For most practical purposes the speed is taken to be 300 000 000 metres per second although a more exact value is 299 792 500 metres per second. Although exceedingly fast, they still take a finite time to travel over a given distance. With modern radio techniques, the time for a signal to propagate over a certain distance needs to be taken into account. Radar for example uses the fact that signals take a certain time to travel to determine the distance of a target. Other applications such as mobile phones also need to take account of the time taken for signals to travel to ensure that the critical timings in the system are not disrupted and that signals do not overlap.

Another major element of a radio wave is its wavelength. This is the distance between a given point on one cycle and the same point on the next cycle as shown. The easiest points to choose are the peaks as these are the easiest to locate. The wavelength was used in the early days of radio or wireless to determine the position of a signal on the dial of a set. Although it is not used for this purpose today, it is nevertheless an important feature of any radio signal or for that matter any electromagnetic wave. The position of a signal on the dial of a radio set or its position within the radio spectrum is now determined by its frequency as this provides a more accurate and convenient method for determining the properties of the signal.


The wavelength is the length from a point on one wave to the identical point on the next
The best point to take is usually the peak

Wavelength of an electromagnetic wave

Finally the frequency of an radio signal or electromagnetic wave is of great importance. This is the number of times a particular point on the wave moves up and down in a given time (normally a second). The unit of frequency is the Hertz and it is equal to one cycle per second. This unit is named after the German scientist who discovered radio waves. The frequencies used in radio are usually very high. Accordingly the prefixes kilo, Mega, and Giga are often seen. 1 kHz is 1000 Hz, 1 MHz is a million Hertz, and 1 GHz is a thousand million Hertz i.e. 1000 MHz. Originally the unit of frequency was not given a name and cycles per second (c/s) were used. Some older books may show these units together with their prefixes: kc/s; Mc/s etc. for higher frequencies.

Frequency to Wavelength Conversion

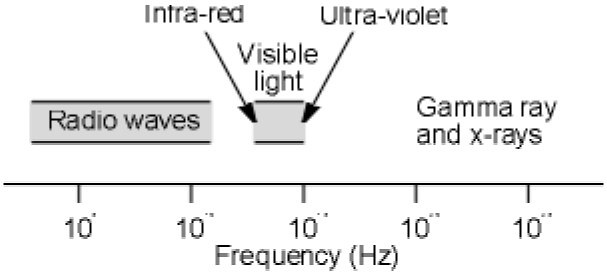
Although wavelength was used as a measure for signals, frequencies are used exclusively today. It is very easy to relate the frequency and wavelength as they are linked by the speed of light as shown:

lambda = c / f

where lambda = the wavelength in metres
f = frequency in Hertz
c = speed of radio waves (light) taken as 300 000 000 metres per second for all practical purposes.

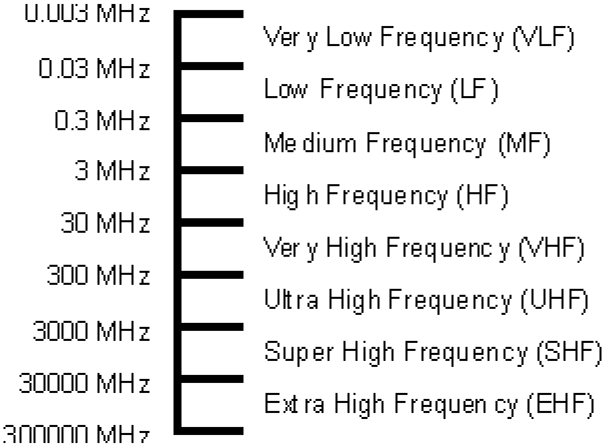
Electromagnetic and the radio spectrum

Electromagnetic waves have an enormous range, and as a result it is very convenient to see where each of the different forms of radiations fits within the spectrum as a whole. It can be seen that radio signals have the lowest frequency, and hence the longest wavelengths. Above the radio spectrum, other forms of radiation can be found. These include infra red radiation, light, ultraviolet and a number of other forms of radiation.



The spectrum of electromagnetic waves

Even within the radio spectrum there is an enormous range of frequencies. It extends over many decades. In order to be able to categorise the different areas and to split the spectrum down into more manageable sizes, the spectrum is split into different segments.



The spectrum of electromagnetic waves

Summary

Electromagnetic waves, and in the context of this website, radio waves, are affected by the media through which they travel. The degree to which they are affected, and the way in which they are affected, often depends upon the frequency. Signals can be reflected, refracted, and diffracted, altering their paths. At other times signals may be absorbed, reducing the distances over which they may travel. However many signals may travel for light years, as in the case of signals travelling from stars. However, for most of us the distances that can be reached by a Wi-Fi or cellular telephone may be of more interest.

Electromagnetic waves - reflection, refraction, diffraction

- a summary or tutorial about the basics of the way in which electromagnetic waves are reflected, refracted and diffracted

This radio path loss and link budget tutorial is split into several pages each of which address different aspects of radio path loss and link budget:

[1] Electromagnetic waves [2] Reflection, refraction, diffraction of e/m waves [3] Polarisation of e/m waves [4] Electromagnetic spectrum

As electromagnetic waves, and in this case, radio signals travel, they interact with objects and the media in which they travel. As they do this the radio signals can be reflected, refracted or diffracted. These interactions cause the radio signals to change direction, and to reach areas which would not be possible if the radio signals travelled in a direct line.

Reflection

Reflection of light is an everyday occurrence. Mirrors are commonplace and can be seen in houses and many other places. Shop windows also provide another illustration for this phenomenon, as do many other areas. Radio waves are similarly reflected by many surfaces.

When reflection occurs, it can be seen that the angle of incidence is equal to the angle of reflection for a conducting surface as would be expected for light. When a signal is reflected there is normally some loss of the signal, either through absorption, or as a result of some of the signal passing into the medium.

A variety of surfaces can reflect radio signals. For long distance communications, the sea provides one of the best reflecting surfaces. Other wet areas provide good reflection of radio signals. Desert areas are poor reflectors and other types of land fall in between these two extremes. In general, though, wet areas provide better reflectors.

For relatively short range communications, many buildings, especially those with metallic surfaces provide excellent reflectors of radio energy. There are also many other metallic structures such as warehouses that give excellent reflecting surfaces. As a result of this signals travelling to and from cellular phones often travel via a variety of paths. Similar effects are noticed for Wi-Fi and other short range wireless communications. An office environment contains many surfaces that reflect radio signals very effectively.

Refraction

It is also possible for radio waves to be refracted. The concept of light waves being refracted is very familiar, especially as it can be easily demonstrated by placing a part of stick or pole in water and leaving the remaining section in air. It is possible to see the apparent change or bend as the stick enters the water. Mirages also demonstrate refraction and a very similar effect can be noticed on hot days when a shimmering effect can be seen when looking along a straight road. Radio waves are affected in the same way. It is found that the direction of an electromagnetic wave changes as it moves from an area of one refractive index to another. The angle of incidence and the angle of refraction are linked by Snell's Law that states:

n1 sin (theta 1) = n2 sin (theta 2)

For radio signals there are comparatively few instances where the signals move abruptly from a region with one refractive index, to a region with another. It is far more common for there to be comparatively gradual change. This causes the direction of the signal to bend rather than undergo an immediate change in direction.

Diffraction

Radio signals may also undergo diffraction. It is found that when signals encounter an obstacle they tend to travel around them. This can mean that a signal may be received from a transmitter even though it may be "shaded" by a large object between them. This is particularly noticeable on some long wave broadcast transmissions. For example the BBC long wave transmitter on 198 kHz is audible in the Scottish glens where other transmissions could not be heard. As a result the long wave transmissions can be heard in many more places than transmissions on VHF FM.

To understand how this happens it is necessary to look at Huygen's Principle. This states that each point on a spherical wave front can be considered as a source of a secondary wave front. Even though there will be a shadow zone immediately behind the obstacle, the signal will diffract around the obstacle and start to fill the void. It is found that diffraction is more pronounced when the obstacle becomes sharper and more like a "knife edge". For a radio signal a mountain ridge may provide a sufficiently sharp edge. A more rounded hill will not produce such a marked effect. It is also found that low frequency signals diffract more markedly than higher frequency ones. It is for this reason that signals on the long wave band are able to provide coverage even in hilly or mountainous terrain where signals at VHF and higher would not.

Polarisation of electromagnetic waves

- and their importance in radio wave propagation

This radio path loss and link budget tutorial is split into several pages each of which address different aspects of radio path loss and link budget:

[1] Electromagnetic waves [2] Reflection, refraction, diffraction of e/m waves [3] Polarisation of e/m waves [4] Electromagnetic spectrum

The polarisation of electromagnetic waves often has a significant effect on the way in which radio wave propagate. While it is important to match the polarisation of the transmitting and receiving antennas, the choice of polarisation is also important for the signal propagation.

What is polarisation

The polarisation of an electromagnetic wave indicates the plane in which it is vibrating. As electromagnetic waves consist of an electric and a magnetic field vibrating at right angles to each other it is necessary to adopt a convention to determine the polarisation of the signal. For this purpose the plane of the electric field is used.

Vertical and horizontal polarisations are the most straightforward forms and they fall into a category known as linear polarisation. Here the wave can be thought of as vibrating in one plane, i.e. up and down, or side to side. This form of polarisation is the most commonly used, and the most straightforward.

However this is not the only form as it is possible to generate waveforms that have circular polarisation. Circular polarisation can be visualised by imagining a signal propagating from an antenna that is rotating. The tip of the electric field vector can be seen to trace out a helix or corkscrew as it travels away from the antenna. Circular polarisation can be either right or left handed dependent upon the direction of rotation as seen from the transmitting antenna.

It is also possible to obtain elliptical polarisation. This occurs when there is a combination of both linear and circular polarisation. Again this can be visualised by imagining the tip of the electric field tracing out an elliptically shaped corkscrew.

Importance for propagation

For many terrestrial applications it is found that once a signal has been transmitted then its polarisation will remain broadly the same. However reflections from objects in the path can change the polarisation. As the received signal is the sum of the direct signal plus a number of reflected signals the overall polarisation of the signal can change slightly although it usually remains broadly the same. When reflections take place from the ionosphere, then greater changes may occur.

In some applications there are performance differences between horizontal and vertical polarisation. For example medium wave broadcast stations generally use vertical polarisation because ground wave propagation over the earth is considerably better using vertical polarisation, whereas horizontal polarisation shows a marginal improvement for long distance communications using the ionosphere. Circular polarisation is sometimes used for satellite communications as there are some advantages in terms of propagation and in overcoming the fading caused if the satellite is changing its orientation.

HF Ionospheric Radio Signal Propagation

- the basics of HF ionospheric radio propagation and how the ionosphere enables radio communications links to be established over large distances around the globe using what are termed sky waves or skywaves.

This radio path loss and link budget tutorial is split into several pages each of which address different aspects of radio path loss and link budget:

[1] HF ionospheric propagation basics [2] The ionosphere [3] Critical frequency, MUF and LUF [4] HF ionospheric propagation and sunspots [5] Ionospheric absorption [6] How the sun affects radio signal propagation [7] Sunspots [8] Solar disturbances [9] SID sudden ionospheric disturbance [10] Solar indices for prediction radio propagation [11] Radio propagation prediction software [12] NVIS near vertical incidence skywave [13] Sporadic E

As electromagnetic waves, and in this case, radio signals travel, they interact with objects and the media in which they travel. As they do this the radio signals can be reflected, refracted or diffracted. These interactions cause the radio signals to change direction, and to reach areas which would not be possible if the radio signals travelled in a direct line.

HF radio communications is dependent for most of its applications on the use of the ionosphere. This region in the atmosphere enables radio communications signals to be reflected, or more correctly refracted back to earth so that they can travel over great distances around the globe. Ionospheric propagation is normally though of as an HF propagation mode, although, it use can extend above and below the HF portion of the spectrum on many occasions.

The fact that radio communications signals can travel all over the globe on the HF bands is widely used by many by broadcasters, news agencies, maritime, radio hams and many other users. Radio transmitters using relatively low powers can be used to communicate to the other side of the globe. Although radio propagation using the ionosphere may not be not as reliable as that provided by satellites, it nevertheless provides a very cost effective and efficient form of radio communication. To enable the most to be made of ionospheric propagation many radio users make extensive use of HF propagation programmes to predict the areas of the globe to which signals may travel, or the probability of them reaching a given area.

These HF propagation prediction programmes utilise a large amount of data, and many have been developed over many years, along with data about the prevailing conditions. However it is still useful to gain a view of how signals travel when using ionospheric propagation and to understand why signal conditions change. In this way the best use can be made of ionospheric propagation.

Radio communications signals in the medium and short wave bands travel by two basic means. The first is known as a ground wave (covered on a separate page in this section), and the second a sky wave using the ionosphere.

Skywaves

When using ionospheric radio propagation, the radio signals leave the Earth's surface and travel towards the ionosphere where some of these are returned to Earth. These radio signals are termed sky waves for obvious reason. If they are returned to Earth, then the ionosphere may (very simply) be viewed as a vast reflecting surface encompassing the Earth that enables signals to travel over much greater distances than would otherwise be possible. Naturally this is a great over simplification because the frequency, time of day and many other parameters govern the reflection, or more correctly the refraction of signals back to Earth. There are in fact a number of layers, or more correctly regions within the ionosphere, and these act in different ways as described below.

D region

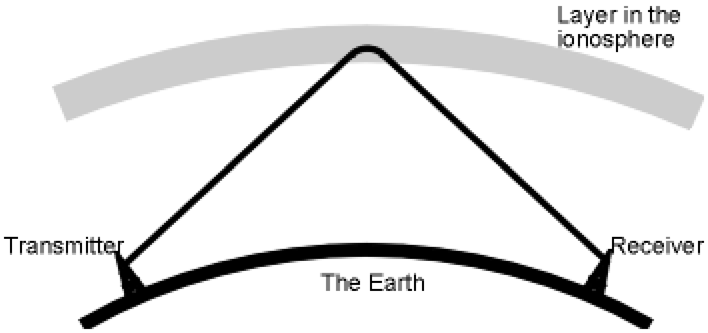
When a sky wave leaves the Earth's surface and travels upwards, the first region of interest that it reaches in the ionosphere is called the D region. This region attenuates the signals as they pass through. The level of attenuation depends on the frequency. Low frequencies are attenuated more than higher ones. In fact it is found that the attenuation varies as the inverse square of the frequency, i.e. doubling the frequency reduces the level of attenuation by a factor of four. This means that low frequency signals are often prevented from reaching the higher regions, except at night when the region disappears.

The D region attenuates signals because the radio signals cause the free electrons in the region to vibrate. As they vibrate the electrons collide with molecules, and at each collision there is a small loss of energy. With countless millions of electrons vibrating, the amount of energy loss becomes noticeable and manifests itself as a reduction in the overall signal level. The amount of signal loss is dependent upon a number of factors: One is the number of gas molecules that are present. The greater the number of gas molecules, the higher the number of collisions and hence the higher the attenuation. The level of ionisation is also very important. The higher the level of ionisation, the greater the number of electrons that vibrate and collide with molecules. The third main factor is the frequency of the signal. As the frequency increases, the wavelength of the vibration shortens, and the number of collisions between the free electrons and gas molecules decreases. As a result signals lower in the radio frequency spectrum are attenuated far more than those which are higher in frequency. Even so high frequency signals still suffer some reduction in signal strength.

E and F Regions

Once a signal passes through the D region, it travels on and reaches first the E, and next the F regions. At the altitude where these regions are found the air density is very much less, and this means that when the free electrons are excited by radio signals and vibrate, far fewer collisions occur. As a result the way in which these regions act is somewhat different. The electrons are again set in motion by the radio signal, but they tend to re-radiate it. As the signal is travelling in an area where the density of electrons is increasing, the further it progresses into the region, the signal is refracted away from the area of higher electron density. In the case of HF signals, this refraction is often sufficient to bend them back to earth. In effect it appears that the region has "reflected" the signal.

The tendency for this "reflection" is dependent upon the frequency and the angle of incidence. As the frequency increases, it is found that the amount of refraction decreases until a frequency is reached where the signals pass through the region and on to the next. Eventually a point is reached where the signal passes through all the regions and on into outer space.



Refraction of a radio signal as it enters an ionised region

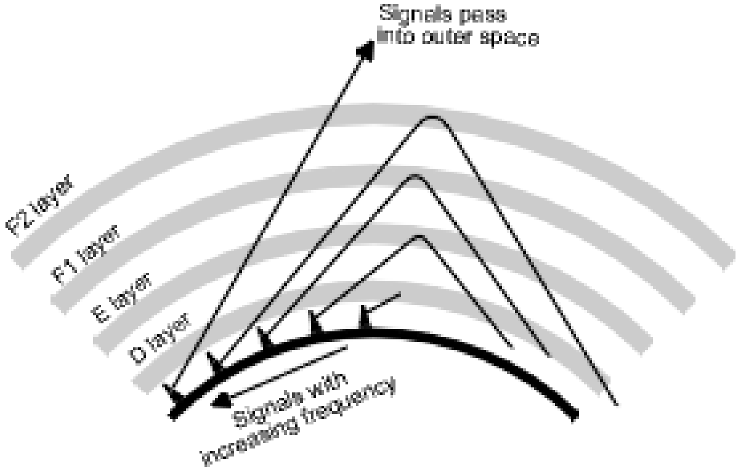
Different frequencies

To gain a better idea of the characteristics of HF propagation using the ionosphere, it is worth viewing what happens to a radio communications signal if the frequency is increased across the frequency spectrum. First it starts with a signal in the medium wave broadcast band. During the day signals on these frequencies only propagate using the ground wave. Any signals that reach the D region are absorbed. However at night as the D region disappears signals reach the other regions and may be heard over much greater distances.

If the frequency of the signal is increased, a point is reached where the signal starts to penetrate the D region and signals reach the E region. Here it is reflected and will pass back through the D region and return to earth a considerable distance away from the transmitter.

As the frequency is increased further the signal is refracted less and less by the E region and eventually it passes right through. It then reaches the F1 region and here it may be reflected passing back through the D and E regions to reach the earth again. As the F1 region is higher than the E region the distance reached will be greater than that for an E region reflection.

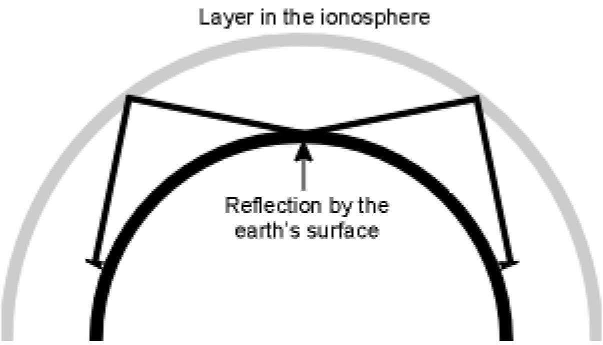
Finally as the frequency of the radio communications signal rises still further the it will eventually pass through the F1 region and onto the F2 region. This is the highest of the regions in the ionosphere and the distances reached using this are the greatest. As a rough guide the maximum skip distance for the E region is around 2500 km and 5000 km for the F2 region.


Signals pass into outer space
F2 layer
F1 layer
E layer
D layer
Signals with increasing frequency

Signals reflected by the E and F regions

Multiple hops

Whilst it is possible to reach considerable distances using the F region as already described, on its own this does not explain the fact that radio signals are regularly heard from opposite sides of the globe using HF propagation with the ionosphere. This occurs because the signals are able to undergo several "reflections". Once the signals are returned to earth from the ionosphere, they are reflected back upwards by the earth's surface, and again they are able to undergo another "reflection" by the ionosphere. Naturally the signal is reduced in strength at each "reflection", and it is also found that different areas of the Earth reflect radio signals differently. As might be anticipated the surface of the sea is a very good reflector, whereas desert areas are very poor. This means that signals that are "reflected" back to the ionosphere by the Pacific or Atlantic oceans will be stronger than those that use the Sahara desert or the red centre of Australia.



Multiple reflections

It is not just the Earth's surface that introduces losses into the signal path. In fact the major cause of loss is the D region, even for frequencies high up into the HF portion of the spectrum. One of the reasons for this is that the signal has to pass through the D region twice for every reflection by the ionosphere. This means that to get the best signal strengths it is necessary signal paths enable the minimum number of hops to be used. This is generally achieved using frequencies close to the maximum frequencies that can support communications using ionospheric propagation, and thereby using the highest regions in the ionosphere. In addition to this the level of attenuation introduced by the D region is also reduced. This means that a radio signal on 20 MHz for example will be stronger than one on 10 MHz if propagation can be supported at both frequencies.

HF propagation summary

HF propagation using the ionosphere is still a widely used as a form of radio communications. While not as reliable as satellite communications, it is not nearly as expensive, and can provide a useful back-up in case the satellite communications fail. It is also widely used as the primary form of radio communications by many organisations from radio broadcasters to radio amateurs, as well as ship to shore and many other forms of point to point communications. As a result HF propagation using the ionosphere is likely to remain in use indefinitely as a form of radio communications technology.

- **Home**
-
- Antennas & propagation
- Broadcast technology
- Cellular telecoms
- Circuits
- Design principles
- Electronic components
- Electronics manufacture
- Formulae & data
- Radio receivers
- RF topics

The Ionosphere and Radiowave Propagation

- an overview or tutorial about the ionosphere, and how it affects radiowave propagation and radio communications.

This radio path loss and link budget tutorial is split into several pages each of which address different aspects of radio path loss and link budget:

[1] HF ionospheric propagation basics [2] The ionosphere [3] Critical frequency, MUF and LUF [4] HF ionospheric propagation and sunspots [5] Ionospheric absorption [6] How the sun affects radio signal propagation [7] Sunspots [8] Solar disturbances [9] SID sudden ionospheric disturbance [10] Solar indices for prediction radio propagation [11] Radio propagation prediction software [12] NVIS near vertical incidence skywave [13] Sporadic E

As electromagnetic waves, and in this case, radio signals travel, they interact with objects and the media in which they travel. As they do this the radio signals can be reflected, refracted or diffracted. These interactions cause the radio signals to change direction, and to reach areas which would not be possible if the radio signals travelled in a direct line.

The ionosphere is a particularly important region with regards to radio signal propagation and radio communications in general. Its properties govern the ways in which radio communications, particularly in the HF radio communications bands take place.

The ionosphere is a region of the upper atmosphere where there are large concentrations of free ions and electrons. While the ions give the ionosphere its name, but it is the free electrons that affect the radio waves and radio communications. In particular the ionosphere is widely known for affecting signals on the short wave radio bands where it "reflects" signals enabling these radio communications signals to be heard over vast distances. Radio stations have long used the properties of the ionosphere to enable them to provide worldwide radio communications coverage. Although today, satellites are widely used, HF radio communications using the ionosphere still plays a major role in providing worldwide radio coverage.

The ionosphere extends over more than one of the meteorological areas, encompassing the mesosphere and the thermosphere, it is an area that is characterised by the existence of positive ions (and more importantly for radio signals free electrons) and it is from the existence of the ions that it gains its name.

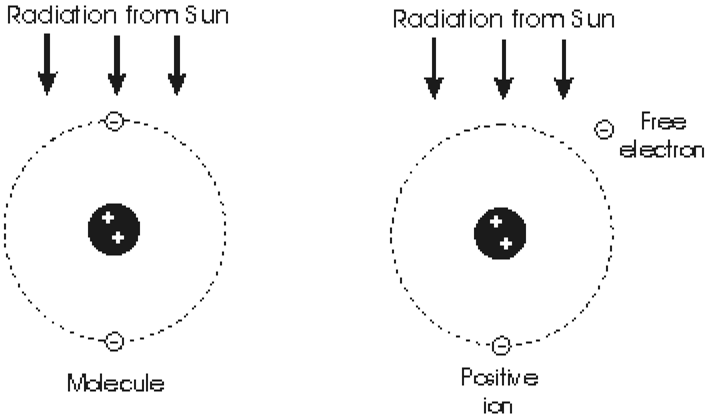
Basics

The free electrons do not appear over the whole of the atmosphere. Instead it is found that the number of free electrons starts to rise at altitudes of approximately 30 kilometres. However it is not until altitudes of around 60 to 90 kilometres are reached that the concentration is sufficiently high to start to have a noticeable effect on radio signals and hence on radio communications systems. It is at this level that the ionosphere can be said to start.

The ionisation in the ionosphere is caused mainly by radiation from the Sun. In addition to this, the very high temperatures and the low pressure result in the gases in the upper reaches of the atmosphere existing mainly in a monatomic form rather than existing as molecules. At lower altitudes, the gases are in the normal molecular form, but as the altitude increases the monatomic forms are more in abundance, and at altitudes of around 150 kilometres, most of the gases are in a monatomic form. This is very important because it is found that the monatomic forms of the gases are very much easier to ionise than the molecular forms.

Ionisation

The Sun emits vast quantities of radiation of all wavelengths and this travels towards the Earth, first reaching the outer areas of the atmosphere. In creating the ionisation it is found that when radiation of sufficient intensity strikes an atom or a molecule, energy may be removed from the radiation and an electron removed, producing a free electron and a positive ion. In the example given below, the simple example of a helium atom is give, although other gases including oxygen and nitrogen are far more common.



Ionisation of molecules by solar radiation

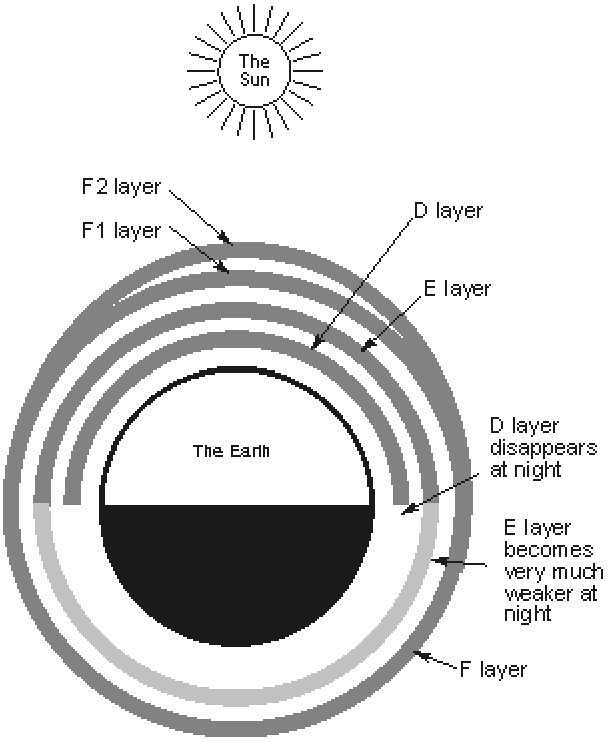
The radiation from the Sun covers a vast spectrum of wavelengths. However in terms of the effect it has on the atoms of molecules it can be considered as photons. The electrons in the atoms or molecules can be considered as orbiting the central nucleus consisting of protons and neutrons. Electrons are tied or bound to their orbit around the nucleus by electro-static forces, the electron is negatively charged and the nucleus is positively charged. There are equal numbers of electrons and protons in any molecule and as a result it is electro-statically neutral.

When a photon strikes the atom, or molecule, the photon transfers its energy to the electron as excess kinetic energy. Under some circumstances this excess energy may exceed the binding energy in the atom or molecule and the electron escapes the influence of the positive charge of the nucleus. This leaves a positively charged nucleus or ions and a negatively charged electron, although as there are the same number of positive ions and negative electrons the whole gas still remains with an overall neutral charge.

Most of the ionisation in the ionosphere results from ultraviolet light, although this does not mean that other wavelengths do not have some effect. Additionally, each time an atom or molecule is ionised a small amount of energy is used. This means that as the radiation passes further into the atmosphere, its intensity reduces. It is for this reason that the ultraviolet radiation causes most of the ionisation in the upper reaches of the ionosphere, but at lower altitudes the radiation that is able to penetrate further cause more of the ionisation. Accordingly, extreme ultra-violet and X-Rays give rise to most of the ionisation at lower altitudes. This reduction in these forms of radiation protects us on the surface of the Earth from the harmful effects of these rays.

The level of ionisation varies over the extent of the ionosphere, being far from constant. One reason is that the level of radiation reduces with decreasing altitude. Also the density of the gases varies. In addition to this there is a variation in the proportions of monatomic and molecular forms of the gases, the monatomic forms of gases being far greater at higher altitudes. These and a variety of other phenomena mean that there are variations in the level of ionisation with altitude.

The level of ionisation in the ionosphere also changes with time. It varies with the time of day, time of year, and according to many other external influences. One of the main reasons why the electron density varies is that the Sun, which gives rise to the ionisation is only visible during the day. While the radiation from the Sun causes the atoms and molecules to split into free electrons and positive ions. The reverse effect also occurs. When a negative electron meets a positive ion, the fact that dissimilar charges attract means that they will be pulled towards one another and they may combine. This means that two opposite effects of splitting and recombination are taking place. This is known as a state of dynamic equilibrium. Accordingly the level of ionisation is dependent upon the rate of ionisation and recombination. This has a significant effect on radio communications.

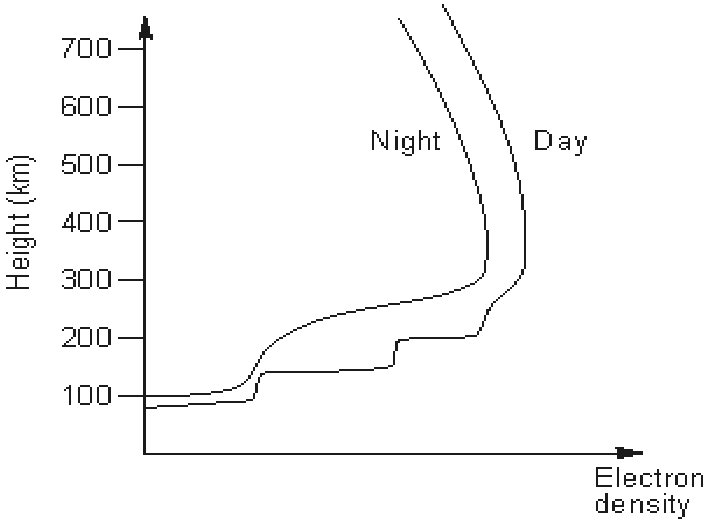

The Sun
F2 layer
F1 layer
D layer
E layer
The Earth
D layer disappears at night
E layer becomes very much weaker at night
F layer

A simplified view of the layers in the ionosphere over the period of a day

Other effects like the season and the state of the Sun also have a major effect. Sunspots and solar disturbances have a major impact on the level of radiation received, and these effects are covered in other articles on this website on Sunspots and Solar Disturbances. The season also has an effect. Again this is covered in other articles on the Radio-Electronics.Com website. However very briefly, the radiation received from the Sun varies in the same way that heat from the Sun varies according to the season, and accordingly the level of ionisation and free electrons changes. However this is a very simplified view as other facts also come into play.

Ionospheric layers

The traditional view of the ionosphere indicates a number of distinct layers, each affecting radio communications in slightly different ways. Indeed, the early discoveries of the ionosphere indicated that a number of layers were present. While this is a convenient way of picturing the structure of the ionosphere it is not exactly correct. Ionisation exists over the whole of the ionosphere, its level varying with altitude. The peaks in level may be considered as the different layers or possibly more correctly, regions. These regions are given letter designations: D, E, and F regions. There is also a C region below the others, but the level of ionisation is so low that it does not have any effect radio signals and radio communications, and it is rarely mentioned.



The typical electron distribution in the ionosphere

The different layers or regions in the ionosphere have different characteristics and affect radio communications in different ways. There are also differences in the exact way they are created and sustained. In view of this it is worth taking a closer look at each one in detail and the way they vary over the complete day during light and darkness.

D Region

The D region is the lowest of the regions within the ionosphere that affects radio communications signals to any degree. It is present at altitudes between about 60 and 90 kilometres and the radiation within it is only present during the day to an extent that affects radio waves noticeably. It is sustained by the radiation from the Sun and levels of ionisation fall rapidly at dusk when the source of radiation is removed. It mainly has the affect of absorbing or attenuating radio communications signals particularly in the LF and MF portions of the radio spectrum, its affect reducing with frequency. At night it has little effect on most radio communications signals although there is still a sufficient level of ionisation for it to refract VLF signals.

The layer is chiefly generated by the action of a form of radiation known as Lyman radiation which has a wavelength of 1215 Angstroms and ionises nitric oxide gas present in the atmosphere. Hard X-Rays also contribute to the ionisation, especially towards the peak of the solar cycle.

E Region

The region above the D region is the E region. It exists at altitudes between about 100 and 125 kilometres. Instead of attenuating radio communications signals this layer chiefly refracts them, often to a degree where they are returned to earth. As such they appear to have been reflected by this layer. However this layer still acts as an attenuator to a certain degree.

Like the D region, the level of ionisation falls relatively quickly after dark as the electrons and ions re-combine and it virtually disappears at night. However the residual night time ionisation in the lower part of the E region causes some attenuation of signals in the lower portions of the HF part of the radio communications spectrum.

The ionisation in this region results from a number of types of radiation. Soft X-Rays produce much of the ionisation, although extreme ultra-violet (EUV) rays (very short wavelength ultra-violet light) also contribute. Broadly the radiation that produces ionisation in this region has wavelengths between about 10 and 100 Angstroms. The degree to which all of the constituents contribute depends upon the state of the Sun and the latitude at which the observations are made.

F Region

The most important region in the ionosphere for long distance HF radio communications is the F region. During the daytime when radiation is being received from the Sun, it often splits into two, the lower one being the F1 region and the higher one, the F2 region. Of these the F1 region is more of an inflection point in the electron density curve (seen above) and it generally only exists in the summer.

Typically the F1 layer is found at around an altitude of 300 kilometres with the F2 layer above it at around 400 kilometres. The combined F layer may then be centred around 250 to 300 kilometres. The altitude of the all the layers in the ionosphere layers varies considerably and the F layer varies the most. As a result the figures given should only be taken as a rough guide. Being the highest of the ionospheric regions it is greatly affected by the state of the Sun as well as other factors including the time of day, the year and so forth.

The F layer acts as a "reflector" of signals in the HF portion of the radio spectrum enabling world wide radio communications to be established. It is the main region associated with HF signal propagation.

Like the D and E layers the level of ionisation of the F region varies over the course of the day, falling at night as the radiation from the Sun disappears. However the level of ionisation remains much higher. The density of the gases is much lower and as a result the recombination of the ions and electrons takes place more slowly, at about a quarter of the rate that it occurs in the E region. As a result of this it still has an affect on radio signals at night being able to return many to Earth, although it has a reduced effect in some aspects.

The F region is at the highest region in the ionosphere and as such it experiences the most solar radiation. Much of the ionisation results from ultra-violet light in the middle of the spectrum as well as those portions of the spectrum with very short wavelengths. Typically the radiation that causes the ionisation is between the wavelengths of 100 and 1000 Angstroms, although extreme ultra-violet light is responsible for some ionisation in the lower areas of the F region.

Summary

The ionosphere is a continually changing area of the atmosphere. Extending from altitudes of around 60 kilometres to more than 400 kilometres it contains ions and free electrons. The free electrons affect the ways in which radio waves propagate in this region and they have a significant effect on HF radio communications.

The ionosphere can be categorised into a number of regions corresponding to peaks in the electron density. These regions are named the D, E, and F regions. In view of the fact that the radiation from the Sun is absorbed as it penetrates the atmosphere, different forms of radiation give rise to the ionisation in the different regions as outlined in the summary table below:

Summary of forms of radiation causing ionisation in the ionosphere.

Region	**Primary Ionising Radiation Forms**
C	Cosmic
D	Lyman alpha, Hard X-Rays
E	Soft X-Rays and some Extreme Ultra-Violet
F1	Extreme Ultra-violet, and some Ultra-Violet
F2	Ultra-Violet

The ionosphere is a continually changing area. It is obviously affected by radiation from the Sun, and this changes as a result aspects including of the time of day, the geographical area of the world, and the state of the Sun. As a result radio communications using the ionosphere change from one day to the next, and even one hour to the next. Predicting how what radio communications will be possible and radio signals may propagate is of great interest to a variety of radio communications users ranging from broadcasters to radio amateurs and two way radio communications systems users to those with maritime mobile radio communications systems and many more.

Critical Frequency, LUF, and MUF

- the Critical Frequency, Lowest Useable Frequency, LUF, and Maximum Usable Frequency, MUF used in radio communications and radio signal propagation

This radio path loss and link budget tutorial is split into several pages each of which address different aspects of radio path loss and link budget:

[1] HF ionospheric propagation basics [2] The ionosphere [3] Critical frequency, MUF and LUF [4] HF ionospheric propagation and sunspots [5] Ionospheric absorption [6] How the sun affects radio signal propagation [7] Sunspots [8] Solar disturbances [9] SID sudden ionospheric disturbance [10] Solar indices for prediction radio propagation [11] Radio propagation prediction software [12] NVIS near vertical incidence skywave [13] Sporadic E

As electromagnetic waves, and in this case, radio signals travel, they interact with objects and the media in which they travel. As they do this the radio signals can be reflected, refracted or diffracted. These interactions cause the radio signals to change direction, and to reach areas which would not be possible if the radio signals travelled in a direct line.

When looking at ionospheric or short wave radio signal propagation for planning a radio communications network or system, or when predicting the HF propagation conditions, there are several frequencies that are important, and are often mentioned in radio signal propagation prediction programmes and in other literature associated with signal propagation. The frequencies include the Critical Frequency, the Lowest Useable Frequency (LUF), and the Maximum Usable Frequency (MUF). Their definitions are at the centre of determining which frequencies will provide the optimum performance for the radio communications system or network.

Critical Frequency

The critical frequency is an important figure that gives an indication of the state of the ionosphere and the resulting HF propagation. It is obtained by sending a signal pulse directly upwards. This is reflected back and can be received by a receiver on the same site as the transmitter. The pulse may be reflected back to earth, and the time measured to give an indication of the height of the layer. As the frequency is increased a point is reached where the signal will pass right through the layer, and on to the next one, or into outer space. The frequency at which this occurs is called the critical frequency.

The equipment used to measure the critical frequency is called an ionosonde. In many respects it resembles a small radar set, but for the HF bands. Using these sets a plot of the reflections against frequency can be generated. This will give an indication of the state of the ionosphere for that area of the world

LUF - Lowest Usable Frequency

As the frequency of a transmission is reduced further reflections from the ionosphere may be needed, and the losses from the D layer increase. These two effects mean that there is a frequency below which radio communications between two stations will be lost. In fact the Lowest Usable Frequency (LUF) is defined as the frequency at below which the signal falls below the minimum strength required for satisfactory reception.

From this it can be seen that the LUF is dependent upon the stations at either end of the path. Their antennas, receivers, transmitter powers, the level of noise in the vicinity, and so forth all affect the LUF. The type of modulation used also has an affect, because some types of modulation can be copied at lower strengths than others. In other words the LUF is the practical limit below which communication cannot be maintained between two particular radio communications stations.

If it is necessary to use a frequency below the LUF then as a rough guide a gain of 10dB must be made to decrease the LUF by 2 MHz. This can be achieved by methods including increasing the transmitter powers, improving the antennas, etc..

MUF - Maximum Usable Frequency

When a signal is transmitted using HF propagation, over a given path there is a maximum frequency that can be used. This results from the fact that as the signal frequency increases it will pass through more layers and eventually travelling into outer space. As it passes through one layer it may be that communication is lost because the signal then propagates over a greater distance than is required. Also when the signal passes through all the layers communication will be lost.

The frequency at which radio communications just starts to fail is known as the Maximum Usable Frequency (MUF). It is generally three to five times the critical frequency, dependent upon the layer being used and the angle of incidence.

Optimum working frequency

To be able to send signals to a given location there are likely to be several different paths that can be used. Sometimes it may be possible to use the either the E or the F layers, and sometimes a signal may be reflected first off one and then the other. In fact the picture is rarely as well defined as it may appear from the textbooks. However it is still possible to choose a frequency from a variety of options to help making contact with a given area.

In general the higher the frequency, the better. This is because the attenuation caused by the D layer is less. Although signals may be able to travel through the D layer they may still suffer significant levels of attenuation. As the attenuation reduces by a facto of four for doubling the frequency in use this shows how significant this can be.

Also by increasing the frequency it is likely that a higher layer in the ionosphere will be used. This may result in fewer reflections being required. As losses are incurred at each reflection and each time the signal passes through the D layer, using a higher frequency obviously helps.

When using the higher frequencies it is necessary to ensure that communications are still reliable. In view of the ever-changing state of the ionosphere a general rule of thumb is to use a frequency that is about 20% below the MUF. This should ensure that the signal remains below the MUF despite the short-term changes. However it should be remembered that the MUF will change significantly according to the time of day, and there fore it will be necessary alter the frequency periodically to take account of this.

HF Radio Propagation and Sunspots

- an overview of the basics of the effects that sunspots have on HF radio propagation and radio communications in general including two way radio communications, maritime mobile radio communications, mobile radio, radio broadcasting and amateur radio communications.

This radio path loss and link budget tutorial is split into several pages each of which address different aspects of radio path loss and link budget:

[1] HF ionospheric propagation basics [2] The ionosphere [3] Critical frequency, MUF and LUF [4] HF ionospheric propagation and sunspots [5] Ionospheric absorption [6] How the sun affects radio signal propagation [7] Sunspots [8] Solar disturbances [9] SID sudden ionospheric disturbance [10] Solar indices for prediction radio propagation [11] Radio propagation prediction software [12] NVIS near vertical incidence skywave [13] Sporadic E

As electromagnetic waves, and in this case, radio signals travel, they interact with objects and the media in which they travel. As they do this the radio signals can be reflected, refracted or diffracted. These interactions cause the radio signals to change direction, and to reach areas which would not be possible if the radio signals travelled in a direct line.

HF radio communications of various forms including two way radio communications, maritime mobile radio communications, radio broadcasting, amateur radio communications, and in fact any form of radio communications that uses the HF bands and ionospheric radio propagation is very dependent upon the state of the ionosphere. The higher the levels of radiation received from the Sun, the greater the levels of ionisation in the ionosphere and in general this brings better propagation conditions for HF radio communications.

It is found that the number of sunspots on the Sun has a considerable effect on the levels of radiation emitted and hence impacting on the ionosphere. In turn this has a marked effect on radio communications of all forms. Sunspots are therefore of great interest to anyone involved in HF radio communications, as it affects the radio propagation conditions so significantly.

What are sunspots?

If the sun is viewed by projecting its image onto a screen, dark areas can be seen from time to time. These can last anything from a few hours right up to several weeks. These spots are cool areas (relatively speaking) on the surface of the sun. The temperature is around only 3000�C against a sizzling 6000�C for the rest of the surface. It is much hotter under the surface reaching temperatures in excess of a million degrees Celsius.

NOTE: ***Under no circumstances should the sun be viewed directly, even though dark glasses. In the past many people have had their sight damaged by doing this.***

These sunspots are areas where there is intense magnetic activity. The fields in these areas are enormous and as a result the surface of the sun is disrupted. In these areas the surface cools dramatically causing a darker region to be perceived.

Around the sunspot there is an area called a plage. This is slightly brighter than the surrounding area and it is a large radiator of cosmic rays, ultra-violet light and X-rays. In fact it results in the overall level of radiation coming from the sun to increase. In turn this increased radiation level from around the sunspots causes the ionosphere to become ionised to a greater extent. This means that higher frequencies can be reflected from the ionosphere.

As sunspots appear in groups, especially the larger ones a sunspot number was devised. This is not the number of sunspots that are observed but a number indicating the level of sunspot activity. The number is very closely related to the actual amount radiation received from the Sun. In this way it is a good measure of solar activity. The daily readings are smoothed mathematically to take out the erratic variations to give the Smoothed Sunspot Number. Sometimes the abbreviation SSN is seen, and it is this smoothed sunspot number that it refers to.

Eleven year cycle

The number of sunspots on the surface of the sun varies with time. At times very few or even none may be visible, whereas at other times the number is very much greater. Although the number varies greatly over short periods of time as the sun rotates, careful analysis using the SSN reveals a longer term trend. It is found that over a period of approximately eleven years over which the sunspots vary. At the peak of this cycle conditions on the bands at the top of the short wave spectrum are very good. Low power stations can be heard over remarkably long distances. At the bottom of the cycle bands around 30 MHz will not usually support normal propagation via the ionosphere.

Sunspots have been observed by the Chinese since before the birth of Christ. However it was not until the mid-eighteenth century that astronomers started to keep records of sunspot numbers. By looking at these over the years it is possible to see the trend since then, and the cycles which have occurred since then. Cycle number 22 officially started in September 1986. It started with a sunspot number of 12 and rose rapidly over the following 33 months to reach a peak of 158. From its peak the sunspot number fell slightly and rose again to give a second, smaller peak before falling to bring the cycle to an end in 1996

Summary

The sunspot activity is of great importance to anyone involved in HF radio communications. Whether two way radio communications, maritime mobile communications, general mobile communications, point to point radio links, amateur radio communications, radio broadcasting or whatever form of radio communications. The level of sunspot activity has an enormous effect on the ionosphere and hence on HF radio propagation conditions. Accordingly even a superficial understanding is advantageous.

onospheric Absorption of Radio Signals

- an overview of absorption or loss of radio signals in the ionosphere, particularly by the D layer - an essential element when planning a radio communications network and predicting HF propagation.

This radio path loss and link budget tutorial is split into several pages each of which address different aspects of radio path loss and link budget:

[1] HF ionospheric propagation basics [2] The ionosphere [3] Critical frequency, MUF and LUF [4] HF ionospheric propagation and sunspots [5] Ionospheric absorption [6] How the sun affects radio signal propagation [7] Sunspots [8] Solar disturbances [9] SID sudden ionospheric disturbance [10] Solar indices for prediction radio propagation [11] Radio propagation prediction software [12] NVIS near vertical incidence skywave [13] Sporadic E

As electromagnetic waves, and in this case, radio signals travel, they interact with objects and the media in which they travel. As they do this the radio signals can be reflected, refracted or diffracted. These interactions cause the radio signals to change direction, and to reach areas which would not be possible if the radio signals travelled in a direct line.

The behaviour of the ionosphere is one of the key areas to consider when planning a radio communications network or system, or when predicting HF propagation conditions.

The ionosphere is usually thought of as an area where radio waves on the short wave bands are refracted or reflected back to Earth. However it is also found that signals are reduced in strength or attenuated as they pass through this area. In fact ionospheric absorption can be one of the major contributors to the reduction in strength of signals.

Most of the attenuation occurs in the D layer. There is some in the E and F layers, but the level is very much less than that experienced in the D layer and it can generally be ignored.

When signals enter the D layer they transfer energy to the electrons and set them in motion, vibrating in line with the radio signal. As the electrons vibrate in this manner they can collide with other molecule, ions, or electrons. Each time a collision occurs a small amount of energy is dissipated, and this is manifested as a loss in the strength of the signal.

The amount of energy that is lost depends primarily upon the number of collisions that take place. In turn this also is dependent upon a number of other factors. The first is the number of other molecules, electrons and ions that are present. In the D layer the density of the air is relatively high, and so there are a large number of other molecules around and the number of collisions is high. The second factor is the frequency of the signal. As the frequency is decreased, so the displacement of the vibrations increases and so does the number of collisions. In fact it is found that the amount of ionospheric absorption that occurs varies inversely as the square of the frequency. In other words if the frequency is doubled, then the attenuation will fall by a factor of four. This is one of the major reasons why when a number of bands or frequencies will support HF propagation between two radio stations, then the highest one will yield the better results. It is also found that the level of attenuation is so high for signals on the medium wave radio broadcast band that during the day when the D layer is present, no signals through it, and signals are only propagated via the ground wave. At night when the D layer disappears, signals are heard from much further afield.

The Sun and HF radio propagation

A summary of the way in which the Sun affects ionospheric propagation (HF radio propagation) for two way radio communications, mobile radio communications, broadcasting, etc.

This radio path loss and link budget tutorial is split into several pages each of which address different aspects of radio path loss and link budget:

[1] HF ionospheric propagation basics [2] The ionosphere [3] Critical frequency, MUF and LUF [4] HF ionospheric propagation and sunspots [5] Ionospheric absorption [6] How the sun affects radio signal propagation [7] Sunspots [8] Solar disturbances [9] SID sudden ionospheric disturbance [10] Solar indices for prediction radio propagation [11] Radio propagation prediction software [12] NVIS near vertical incidence skywave [13] Sporadic E

As electromagnetic waves, and in this case, radio signals travel, they interact with objects and the media in which they travel. As they do this the radio signals can be reflected, refracted or diffracted. These interactions cause the radio signals to change direction, and to reach areas which would not be possible if the radio signals travelled in a direct line.

The Sun has an enormous impact on HF radio propagation because it affects the ionosphere which gives rise to most of the long distance effects that enable long distance radio communications on the HF bands. As a result is greatly affects many forms of radio communications from the typically two way radio communications systems that are sued by many organisations, and various forms of mobile radio communications using the HF bands to radio broadcasting, point to point radio communications and radio amateur transmissions. As a result, a knowledge of how the conditions on the Sun affect radio signal propagation is essential for radio planning and prediction HF propagation conditions. Radio propagation prediction software also takes the state of the Sun into consideration when it calculates its estimates of the propagation conditions.

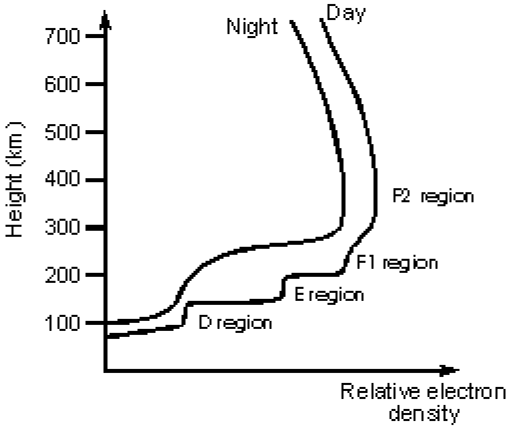
To look at the way the Sun affects the ionosphere and radio propagation conditions, it is necessary to take a quick look at the various areas in the atmosphere to see which areas influence radio propagation and how the sun affects them. These factors are important in being able to predict propagation conditions and when using radio propagation prediction programmes.

The ionosphere

For many years it has been known that there are ionised layers in the upper reaches of the atmosphere that affect various forms of radio communications. This region is known as the ionosphere, although the existence of the an ionised region was first proposed just after the turn of the century, separately by two scientists, namely Kennelly in the USA and Heaviside in the UK. Since then far more has been learned about these layers, especially since the first rockets managed to pass through the ionosphere to collect data.

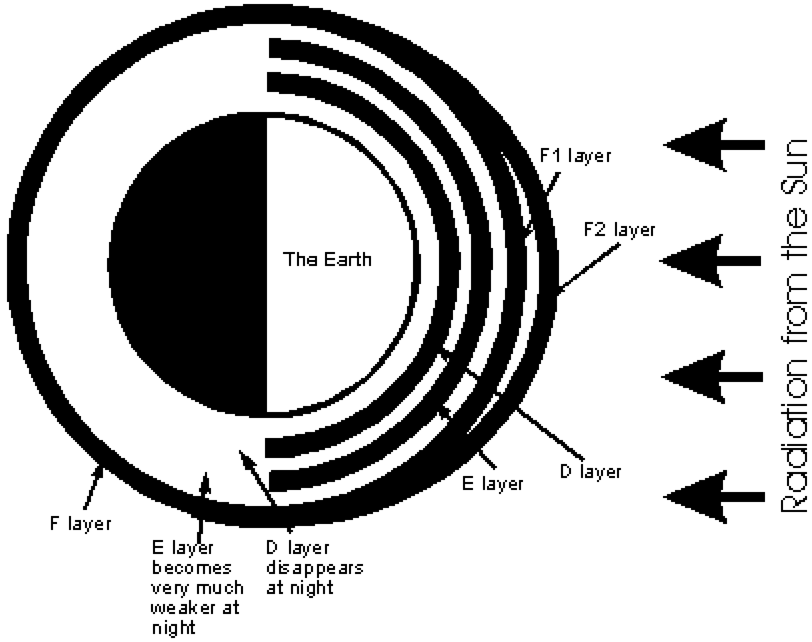
In most regions of the atmosphere it is found that the gases are in a stable molecular form. However in certain areas of the atmosphere some of them start to become ionised, splitting into free electrons and positive ions. Of these it is the free electrons that affect the radio signals, although the layer where these ions and electrons are found is still called the ionosphere. This generally starts to happen at an altitude of around 30 km, although at this height the levels of ionisation are very small and they do not have an effect on radio signals. However as the altitude increases the number of ions rises.

The ionosphere is traditionally thought of as having a number of distinct layers. While it is often convenient to think of the ionosphere in this way, it is not strictly true. The whole of the ionosphere contains ions and free electrons, although there are a number of peaks, and which may be considered as the different layers. These layers are given the designations D, E, and F. A diagram of the approximate levels of ionisation is shown below. This can only be very approximate because the levels of ionisation vary as a result of a number of factors.



The approximate levels of ionisation in the ionosphere

The lowest of the layers is the D layer. This is found at altitudes between 60 and 90 km. It only exists during the daytime when it is in view of the sun. Above this is the E layer at around 110 km. This exists during the day, and then at night when it is not in sunlight it becomes very much weaker. Finally there is the F layer. This varies considerably, normally existing as two layers during the day. These are designated the F1 and F2 layers. They are found at altitudes of around 300 and 400 km in summer, and then during the winter they may fall to around 200 and 300 km. At night the two layers generally combine to form a single layer and this is generally around an altitude of 250 to 300 km. It should be remembered that these figures are only a rough guide because they change quite considerably according to the time of year and the state of the sun.


F1 layer
F2 layer
The Earth
Radiation from the Sun
D layer
E layer
F layer
E layer becomes very much weaker at night
D layer disappears at night

Variations of the ionosphere over the day

Formation of ions

The ionisation in the ionosphere is generated when radiation from the sun strikes the gas molecules in the upper atmosphere. The radiation is of sufficient intensity that it gives the electron in some molecules sufficient energy to leave the molecular structure. This leaves a free electron and the gas molecule, having one electron too few becomes a positive ion.

At very high altitudes the atmosphere is very thin, and as a result the levels of ionisation are very low. As the atmosphere become denser, so the level of ionisation starts to rise. However the ionisation process uses up the energy of the radiation, and after a certain distance the energy of the radiation is such that it does not ionise as many gas molecules as before and the level of ionisation begins to fall.

It is also found that for the higher layers including the F and E layers most of the ionisation results from ultra violet light. The D layer being at a lower altitude results mainly from X-rays that are able to penetrate further into the atmosphere.

It is also found that the free electrons and positive ions slowly recombine. In other words the radiation is causing them to ionise, and then they slowly recombine afterwards. In chemistry this state of affairs is called a dynamic equilibrium. It means that if the source of radiation is removed, then the levels of ionisation will fall. As a result the D layer disappears after nightfall, and the E layer is greatly reduced in intensity. In view of the high levels of ionisation in the F layers and the fact that the air density is so much less, it takes longer for the recombination process to take place and consequently it remains over night, although its level is reduced. This can be seen in the way that radio communications vary over the course of a day.

Effect of the ionosphere

The different layers of the ionosphere affect radio ways in slightly different ways. When a signal enters the D layer it sets the free electrons vibrating. As they vibrate they collide with nearby molecules, and after each collision some energy is lost. As a result radio signals entering the D layer are attenuated. It is found that the level of attenuation is inversely proportional to the square of the frequency. In other words doubling the frequency reduces the attenuation by a factor of four. It is found that low frequency radio signals are completely absorbed by it. This can be shown by the fact that radio stations on the medium wave broadcast band can only be heard over short distances during the day, and then at night when the D layer disappears they can be heard over much greater distances.

The effect is slightly different for the higher layers. Being higher in altitude the gas density is much less. As a result a different effect predominates. Again the electrons are set in motion, but as fewer collisions take place they act on the signal to bend it away from the area of highest ionisation. In other words the signal is refracted back towards the earth. It is also found that the effect decreases with frequency and as a result the signal will eventually pass through one layer and on to the next.

Variations in the ionosphere

The effect of the ionosphere is greatly linked to the amount of radiation it receives. This varies over the period of a day. At night when the ionosphere receives no radiation from the sun, the level of ionisation falls and communication may not be possible over some paths or different frequencies may have to be used.

Other changes also affect the ionosphere. In just the same way that winters are colder because that part of the earth receives less warm from the sun, so the ionosphere receives less radiation, and the levels of ionisation in the ionosphere fall.

Sunspots

Changes on the sun itself also affect the ionosphere. One of the major changes occurs as the result of the sunspots that appear on the surface of the sun.

If the sun is viewed by projecting its image onto a screen, then a number of dark areas may be seen from time to time. These spots may last from anywhere between a few hours to days or even weeks. The spots are areas where the surface of the sun is cooler than the surrounding areas. The temperature of the spots is only about 3000 C. This is quite cool when compared to the temperature of the rest of the surface that is around 6000 C! However it is very much hotter under the surface where temperatures reach in excess of a million degrees.

The sunspots are areas of intense magnetic activity. The magnetic fields in these areas are enormous and as a result the surface of the sun is disrupted. This causes the surface temperature to fall in these areas causing a darker area to be perceived.

Around the sunspot itself there is an area that is known as a plage. This is slightly brighter than the surrounding area and is a large radiator of ultra-violet radiation and X-rays. The amount of radiation emanating from the plage means that there is an overall increase in the level of radiation from the sun. In fact it is noticed that the level of radiation from the sun can be estimated from a knowledge of the number of sunspots that appear on the surface.

As sunspots often appear in groups, a method of trying to estimate their effect has been devised. A figure known as the sunspot number is used. This number does not represent the number of spots themselves, but the level of activity on the sun and the sunspot number is very closely related to the amount of radiation received from the sun.

The daily readings of the sunspot numbers fluctuate considerably. To overcome this, the readings are smoothed mathematically to take out the erratic nature of the readings and so that the underlying trend can be seen. This number, called the Smoothed Sunspot Number (SSN) is often quoted in association with propagation reports.

The sunspot cycle

The number of sunspots on the Sun's surface varies. On some days very few, or even none may be seen, whereas at other times there are very many. The daily number varies considerably over a short period of time as the sun rotates, but if the smoothed sunspot number is used it can be seen that there is a much longer-term trend. This trend shows that the number of sunspots rises and falls over a period of approximately eleven years. This number is only an approximate guide because there is a considerable amount of variation on this.

Records of the sunspot numbers have been kept since the mid-eighteenth century, and by referring to these records it has been possible to track the cycles since then. Cycle 22 started in September 1986 with a number of 12. It rose rapidly over the next 33 months to reach a peak of 158. From then it fell slightly and rose again to give a second smaller peak before ending in 1996. Now cycle 23 has started and the numbers are rising.

The effect of the sunspot cycle

Most short wave radio listeners and DX'ers look upon rising sunspot numbers with anticipation. The increased numbers of sunspots mean increased levels of radiation. In turn this means that there are greater levels of ionisation in the ionosphere. Accordingly this affects propagation on the HF bands. It is found that the maximum frequencies that can be reflected are increased.

At the sunspot minimum frequencies of around 15 to 20 MHz are normally supported during the day. However at the maximum, frequencies in excess of 60MHz may be affected. This means that popular ham bands like 24 and 28 MHz may not support communications via normal ionosphere modes in the sunspot minima. Often 28 MHz appears dead with no stations audible. However during periods of around the maximum it is an excellent band. Low power stations or those with poorer antennas find it particularly good. As the D layer attenuation is much less, even low power stations can make excellent contacts.

The sunspot number can be used to give a very rough guide to what conditions may be like. The figure tends to vary from about 65 at the minimum of the cycle to over 300 at the maximum. For good conditions on the higher frequency bands it is found that a figure of in excess of about 100 is required. Up to date figures can be accessed from a variety of web sites including http://www.sunspotcycle.com

HF Radio Propagation and Sunspots

- an overview of the basics of the effects that sunspots have on HF radio propagation and radio communications in general including two way radio communications, maritime mobile radio communications, mobile radio, radio broadcasting and amateur radio communications.

This radio path loss and link budget tutorial is split into several pages each of which address different aspects of radio path loss and link budget:

[1] HF ionospheric propagation basics [2] The ionosphere [3] Critical frequency, MUF and LUF [4] HF ionospheric propagation and sunspots [5] Ionospheric absorption [6] How the sun affects radio signal propagation [7] Sunspots [8] Solar disturbances [9] SID sudden ionospheric disturbance [10] Solar indices for prediction radio propagation [11] Radio propagation prediction software [12] NVIS near vertical incidence skywave [13] Sporadic E

As electromagnetic waves, and in this case, radio signals travel, they interact with objects and the media in which they travel. As they do this the radio signals can be reflected, refracted or diffracted. These interactions cause the radio signals to change direction, and to reach areas which would not be possible if the radio signals travelled in a direct line.

HF radio communications of various forms including two way radio communications, maritime mobile radio communications, radio broadcasting, amateur radio communications, and in fact any form of radio communications that uses the HF bands and ionospheric radio propagation is very dependent upon the state of the ionosphere. The higher the levels of radiation received from the Sun, the greater the levels of ionisation in the ionosphere and in general this brings better propagation conditions for HF radio communications.

It is found that the number of sunspots on the Sun has a considerable effect on the levels of radiation emitted and hence impacting on the ionosphere. In turn this has a marked effect on radio communications of all forms. Sunspots are therefore of great interest to anyone involved in HF radio communications, as it affects the radio propagation conditions so significantly.

What are sunspots?

If the sun is viewed by projecting its image onto a screen, dark areas can be seen from time to time. These can last anything from a few hours right up to several weeks. These spots are cool areas (relatively speaking) on the surface of the sun. The temperature is around only 3000�C against a sizzling 6000�C for the rest of the surface. It is much hotter under the surface reaching temperatures in excess of a million degrees Celsius.

NOTE: ***Under no circumstances should the sun be viewed directly, even though dark glasses. In the past many people have had their sight damaged by doing this.***

These sunspots are areas where there is intense magnetic activity. The fields in these areas are enormous and as a result the surface of the sun is disrupted. In these areas the surface cools dramatically causing a darker region to be perceived.

Around the sunspot there is an area called a plage. This is slightly brighter than the surrounding area and it is a large radiator of cosmic rays, ultra-violet light and X-rays. In fact it results in the overall level of radiation coming from the sun to increase. In turn this increased radiation level from around the sunspots causes the ionosphere to become ionised to a greater extent. This means that higher frequencies can be reflected from the ionosphere.

As sunspots appear in groups, especially the larger ones a sunspot number was devised. This is not the number of sunspots that are observed but a number indicating the level of sunspot activity. The number is very closely related to the actual amount radiation received from the Sun. In this way it is a good measure of solar activity. The daily readings are smoothed mathematically to take out the erratic variations to give the Smoothed Sunspot Number. Sometimes the abbreviation SSN is seen, and it is this smoothed sunspot number that it refers to.

Eleven year cycle

The number of sunspots on the surface of the sun varies with time. At times very few or even none may be visible, whereas at other times the number is very much greater. Although the number varies greatly over short periods of time as the sun rotates, careful analysis using the SSN reveals a longer term trend. It is found that over a period of approximately eleven years over which the sunspots vary. At the peak of this cycle conditions on the bands at the top of the short wave spectrum are very good. Low power stations can be heard over remarkably long distances. At the bottom of the cycle bands around 30 MHz will not usually support normal propagation via the ionosphere.

Sunspots have been observed by the Chinese since before the birth of Christ. However it was not until the mid-eighteenth century that astronomers started to keep records of sunspot numbers. By looking at these over the years it is possible to see the trend since then, and the cycles which have occurred since then. Cycle number 22 officially started in September 1986. It started with a sunspot number of 12 and rose rapidly over the following 33 months to reach a peak of 158. From its peak the sunspot number fell slightly and rose again to give a second, smaller peak before falling to bring the cycle to an end in 1996

Summary

The sunspot activity is of great importance to anyone involved in HF radio communications. Whether two way radio communications, maritime mobile communications, general mobile communications, point to point radio links, amateur radio communications, radio broadcasting or whatever form of radio communications. The level of sunspot activity has an enormous effect on the ionosphere and hence on HF radio propagation conditions. Accordingly even a superficial understanding is advantageous.

Solar Flares and Disturbances for Radio Propagation

- an overview of the basics of solar flares and other disturbances including CMEs and how they affect ionospheric HF radio propagation conditions for two way radio communications, maritime mobile radio communications, point to point radio communications and radio broadcasting.

This radio path loss and link budget tutorial is split into several pages each of which address different aspects of radio path loss and link budget:

[1] HF ionospheric propagation basics [2] The ionosphere [3] Critical frequency, MUF and LUF [4] HF ionospheric propagation and sunspots [5] Ionospheric absorption [6] How the sun affects radio signal propagation [7] Sunspots [8] Solar disturbances [9] SID

sudden ionospheric disturbance [10] Solar indices for prediction radio propagation [11] Radio propagation prediction software [12] NVIS near vertical incidence skywave [13] Sporadic E

As electromagnetic waves, and in this case, radio signals travel, they interact with objects and the media in which they travel. As they do this the radio signals can be reflected, refracted or diffracted. These interactions cause the radio signals to change direction, and to reach areas which would not be possible if the radio signals travelled in a direct line.

The condition of the Sun has a major impact on ionospheric radio propagation. Accordingly it affects a variety of forms of HF radio communications including two way radio communications, maritime mobile radio communications, general mobile radio communications using the HF bands, point to point radio communications, radio broadcasting and amateur radio communications.

As the Sun provides the radiation that governs the state of the ionosphere and hence HF radio propagation, any flares or other disturbances are of great importance. Under some circumstances these can enhance radio communications and the HF radio propagation conditions. Under other circumstances they can disrupt radio communications on the HF bands, while at the same time providing some radio propagation conditions that can be used at VHF by radio amateurs.

there are a number of types of disturbance that are of particular interest for radio communications. Flares are one of the most obvious. However, apart from solar flares there are other disturbances that occur. One is the coronal mass ejection, and there are also coronal holes.

Solar flares

Solar flares are enormous explosions that occur on the surface of the Sun. They result in the emission of colossal mounts of energy. In addition to this, the larger solar flares also eject large amounts of material mainly in the form of protons.

Flares erupt in just a few minutes with apparently no warning. When they occur the material is heated to millions of degrees Celsius and it leaves the surface of the Sun in a huge arch, returning some time later. The flares normally occur near sunspots, often along the dividing line between them where there are oppositely directed magnetic forces.

It is the magnetic fields appear to be responsible for the solar flares. When the magnetic field between the sunspots becomes twisted and sheared the magnetic field lines may cross and reconnect with enormous explosive energy. When this occurs an eruption of gases takes place through the solar surface, and it extends several tens of thousands of miles out from the surface of the Sun and follow the magnetic lines of force to form a solar flare. The gases from within the sun start to rise and the area becomes heated even more and this causes the level of visible radiation and other forms of radiation to increase.

Solar flare
Image courtesy NASA

During the first stages of the solar flare, high velocity protons are ejected. These travel at around a third the speed of light. Then, about five minutes into the solar flare, lower energy particles follow. This material follows the arc of the magnetic lines of force and returns to the Sun, although some material is ejected into outer space especially during the larger flares.

Effect of solar flares: For most solar flares, the main effect felt on Earth is an increase in the level of solar radiation. This radiation covers the whole electromagnetic spectrum and elements such as the ultra-violet, X-rays and the like will affect the levels of ionisation in the ionosphere and hence it has an effect on radio communications via the ionosphere. Often an enhancement in ionospheric HF propagation is noticed as the higher layers of the ionosphere have increased levels of iononisation. However if the levels of ionisation in the lower elvels start to rise then this can result in higher levels of attenuation of the radio communications signals and poor conditions may be experienced. Additionally an increase in the level of background noise at VHF can also be detected easily.

Flares generally only last for about an hour, after which the surface of the Sun returns to normal although some Post Flare Loops remain for some time afterwards. The flares affect radio propagation and radio communications on Earth and the effects may be noticed for some time afterwards.

Solar Flare Classifications: Flares are classified by their intensity at X-ray wavelengths, i.e. wavelengths between 1 - 8 Angstroms. The X-Ray intensity from the Sun is continually monitored by the National Oceanic and Atmospheric Administration (NOAA) using detectors on some of its satellites. Using this data it is possible to classify the flares. The largest flares are termed X-Class flares. M-Class flares are smaller, having a tenth the X-Ray intensity of the X-Class ones. C-Class flares then have a tenth the intensity of the M-Class ones.

It is found that the occurrence of these flares correlate well with the sunspot cycle, increasing in number towards the peak of the sunspot cycle.

CMEs

Coronal mass ejections, CMEs, are another form of disturbance that can affect radio communications. Although much greater than flares in many respects, CMEs were not discovered until spacecraft could observe the Sun from space. The reason for this is that Coronal Mass Ejections, CMEs can only be viewed by looking at the corona of the Sun, and until the space age this could only be achieved during an eclipse. As eclipses occur very infrequently and only last for a few minutes. Using a space craft the corona could be seen when viewing through a coronagraph, a specialised telescope with what is termed an occulting disk enabling it to cut out the main area of the Sun and only view the corona. This enabled the corona to be viewed.

Although ground based coronagraphs are available, they are only able to view the very bright innermost area of the corona. Space based ones are able to gain a very much better view of the corona extending out to very large distances from the Sun and in this way see far more of the activity in this region, and hence view CMEs.

Coronal Mass Ejections, CMEs are huge bubbles of gas that are threaded with magnetic field lines, and the bubbles are ejected over the space of several hours. For many years it was thought that solar flares were responsible for ejecting the masses of particles that gave rise to the auroral disturbances that are experienced on earth. Now it is understood that CMEs are the primary cause.

It is now understood that CMEs disrupt the steady flow of the solar wind producing a large increase in the flow. This may result in large disturbances that might strike the Earth if they leave the Sun in the direction of the Earth.

Coronal Mass Ejections, CMEs are often associated with solar flares eruptions but they can also occur on their own. Like solar flares their frequency varies according to the position in the sunspot cycle, peaking around the sunspot maximum, and falling around the minimum. At solar minimum there may be about one each week whilst at the peak two or three may be observed each day. Fortunately they do not all affect the Earth. Material is thrown out from the Sun in one general direction and only if this is on an intersecting trajectory will it affect the Earth.

CMEs can give rise to ionospheric storms. These can provide a short lived enhancement to ionospheric radio propagation conditions but before long this can result in a black out to radio communications via the ionosphere.

Coronal Holes

Coronal holes are another important feature of solar activity. They are regions where the corona appears dark. They were first discovered after X-ray telescopes were first launched into space and being above the Earth's atmosphere they were able to study the structure of the corona across the solar disc. Coronal holes are associated with "open" magnetic field lines and are often although not exclusively found at the Sun's poles. The high-speed solar wind is known to originate from them and this has an impact on ionospheric radio propagation conditions and hence on all HF radio communications.

Summary

Solar disturbances are responsible for many of the major changes in the ionosphere. The effects of both CMEs and solar flares can cause major changes to ionospheric radio propagation, often disrupting them for hours or sometimes days. As a result a knowledge of when they are happening, and their size can help in predicting what ionospheric radio conditions may be like.

Sudden Ionospheric Disturbance (SID)

- overview of the sudden ionospheric disturbance, SID, that affects HF radio signal propagation

This radio path loss and link budget tutorial is split into several pages each of which address different aspects of radio path loss and link budget:

[1] HF ionospheric propagation basics [2] The ionosphere [3] Critical frequency, MUF and LUF [4] HF ionospheric propagation and sunspots [5] Ionospheric absorption [6] How the sun affects radio signal propagation [7] Sunspots [8] Solar disturbances [9] SID sudden ionospheric disturbance [10] Solar indices for prediction radio propagation [11] Radio propagation prediction software [12] NVIS near vertical incidence skywave [13] Sporadic E

As electromagnetic waves, and in this case, radio signals travel, they interact with objects and the media in which they travel. As they do this the radio signals can be reflected, refracted or diffracted. These interactions cause the radio signals to change direction, and to reach areas which would not be possible if the radio signals travelled in a direct line.

Sudden Ionospheric Disturbances or SIDs are often experienced on the HF or short wave bands where they may also be called blackouts, short wave fades (SWF), or they are sometimes referred to as the Dellinger effect, after the person who first noted them. As the name implies they occur suddenly, and they cause a blackout of ionospheric radio propagation for a number of hours. This may cover part, or even all of the HF radio spectrum.

Cause

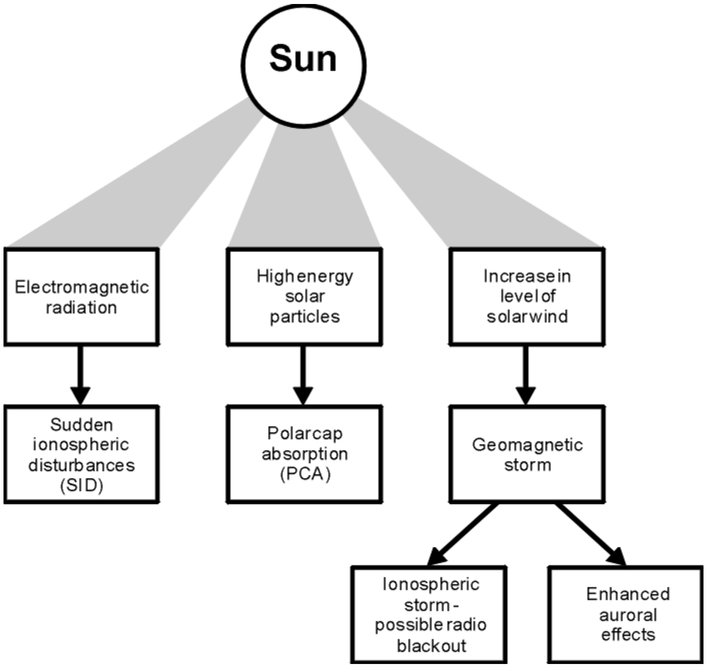
There are a number of types of blackout that may be experienced on the HF bands. Sudden Ionospheric Disturbances, SID, only last for a few hours, but they may be the precursor for a longer blackout. Generally a SID is caused by a large solar flare (M or X class - see the article on Solar Disturbances on this website). Along with the flare there is a massive increase in the level of radiation that is emitted by the Sun. It takes just over 8 minutes for the radiation from the Sun to reach the Earth, at which point the effects start to be noticed and the SID begins. As light takes the same time to reach the Earth, there is no prior warning of an event happening.

The radiation from the Sun caused by the flare is not limited to one form of radiation, but it includes all forms including a high level of X-rays. The X-rays are able to penetrate through to the D layer or D region of the ionosphere and as a result they give rise to a high level of ionisation in the D region. This results in a very significant increase in the level of D region attenuation.

Although the increase in the level of radiation is rapid, it takes time for the level of ionisation to rise. As a result the lower frequencies are affected first, and as the degree of ionisation increases, so the higher frequencies are affected. During a normal day the D region normally affects frequencies up to two or three MHz, but the increase in radiation from a flare can cause much higher frequencies to be affected. Often it can result in a complete black out of the HF bands. At other times a SID may only affect the lower frequencies.

Other effects

There are a number of other side effects that may be noticed, especially at the onset of a Sudden Ionospheric Disturbance, SID. The D region is affected by the large increase in X-rays, but other forms of radiation, including ultra-violet also increase in intensity. This gives rise to an increase in the degree of ionisation in the higher regions of the ionosphere. This can result in an increase in the maximum usable frequency, especially at the beginning of a radio blackout when the D region attenuation is not as high.


Sun
Electromagnetic radiation
High energy solar particles
Increase in level of solar wind
Sudden ionospheric disturbances (SID)
Polar cap absorption (PCA)
Geomagnetic storm
Ionospheric storm - possible radio blackout
Enhanced auroral effects

Summary of the effects of solar disturbances

SID duration

A SID may last only a few hours. As recovery starts, the high frequencies are affected first. After the flare as the level of radiation falls, so will the D region ionisation. As its intensity falls, so it will affect the higher frequencies less, and HF propagation conditions will slowly return to normal unless they are part of a larger disturbance.

It is also worth noting that it is only the sunlit side of the earth that is affected. Any areas that were in darkness when the flare occurred escape the effects. A further effect of a flare is that at the same time as the X-ray radiation reaches the Earth giving rise to a blackout, it is found that noise bursts from the increased levels of radio frequency radiation may be monitored at VHF.

Radio Propagation Prediction using Solar Indices

- an overview, article, tutorial about the basics of radio propagation prediction using solar indices. This is applicable for two way radio communication, mobile radio communications, radio broadcasting, maritime mobile communications and for radio amateurs.

This radio path loss and link budget tutorial is split into several pages each of which address different aspects of radio path loss and link budget:

[1] HF ionospheric propagation basics [2] The ionosphere [3] Critical frequency, MUF and LUF [4] HF ionospheric propagation and sunspots [5] Ionospheric absorption [6] How the sun affects radio signal propagation [7] Sunspots [8] Solar disturbances [9] SID sudden ionospheric disturbance [10] Solar indices for prediction radio propagation [11] Radio propagation prediction software [12] NVIS near vertical incidence skywave [13] Sporadic E

As electromagnetic waves, and in this case, radio signals travel, they interact with objects and the media in which they travel. As they do this the radio signals can be reflected, refracted or diffracted. These interactions cause the radio signals to change direction, and to reach areas which would not be possible if the radio signals travelled in a direct line.

Ionospheric radio propagation is notoriously changeable. However for many applications radio propagation prediction is necessary. For example for broadcasting as well as for users of two way radio communications links that utilise the HF bands as well as mobile radio communications, maritime radio communications, and many other point to point radio users, a knowledge of the propagation patterns that will be in existence at a particular time are almost essential. In this way radio communications users who require propagation via the ionosphere can choose the best times and frequencies in which to establish their radio communications.

Radio propagation prediction

There are many indicators that enable the HF radio propagation conditions to be predicted. However it is indicators of the level of solar radiation and geomagnetic activity that give the best clues to the possible state of radio communications propagation conditions via the ionosphere. The main indicators are the solar flux and the geomagnetic indices. Using these it is possible to manually deduce what conditions may be like. However there several packages of radio propagation prediction software that are available. These take the various indices into account along with the position on the globe, time of day, season, and the position in the sunspot cycle.

Solar flux

One of the major indicators of solar activity used for radio propagation prediction is known as the solar flux and it has a major impact on radio communications propagation conditions. It provides an indication of the level of radiation that is being received from the Sun. This solar index is measured by detecting the level of radio noise emitted at a frequency of 2800 MHz (10.7 cms). The index is quoted in terms of Solar Flux Units (SFU). An SFU has the units 10^-22 Watts per metre^2 per Hz.

The level of solar radiation varies around the globe. Even when correction factors have been applied it is not easy to be able to provide a consist series of figures. To overcome this, the standard is taken as the reading from the Penticton Radio Observatory in British Columbia, Canada. Thus these figures are of great interest for ionospheric radio propagation prediction.

The level of ionising radiation that is received from the Sun is approximately proportional to the Solar Flux. There is a statistical relationship rather than a direct one because the level of radio noise received at 2800 MHz is about a million times less in intensity than that of the radiation that creates the ionisation in the ionosphere. However the Solar Flux provides a good first order approximation, particularly for the F region that is responsible for most long distance ionospheric radio communications propagation. The best correlation is with levels of the Smoothed Sunspot Number (SSN).

It is possible to relate the daily sunspot number to the Solar Flux. A number of equations are available but the one given below is straightforward and sufficiently accurate for most purposes:

$$\textbf{Solar flux (SFU)} = 73.4 + 0.62R$$

Where R is the daily sunspot or Wolf number.

A slightly more accurate, although more complicated equation indicates the relationship between the two values is not totally linear.

$$\textbf{Solar flux (SFU)} = 63.7 + 0.728 R + 0.00089 R^2$$

The values of Solar Flux vary over a wide range. At their lowest (typically during the periods of the sunspot minima) they may be as low as 50 but rise to maximum values of around 300 (around the times of the sunspot maxima).

As the values of Solar Flux provide an indication of the level of ionisation in the ionosphere. In turn this gives an indication of what the Maximum Usable Frequency (MUF) for radio communications may be. Low values of Solar Flux indicate that MUF figures may be low. High values of Solar Flux indicate that the MUF may be higher. It should be remembered that there must be several consecutive days of sustained high solar radiation with the absence of solar disturbances for high MUFs for HF band radio communications.

Geomagnetic indices

Apart from the Solar Flux, another important influence on the ionosphere and hence radio propagation prediction is the level of geomagnetic activity. While the geomagnetic activity is a measure of the state of the Earth's magnetic field, this in turn is influenced by the Sun. To indicate the state of Geomagnetic activity, there are two indices that are used that are related to each other:

- K index
- a Index

Although different, both these indices give indications of the severity of magnetic fluctuations, and hence the level of disturbance to the ionosphere.

K Index: The K index is a three hourly measurement of the variation of the Earth's magnetic field compared to what are "quiet day" conditions. The measurement is made using a magnetometer. This indicates the variation of the magnetic flux in nanoTeslas. This reading is then converted to the K index. The relationship is quasi-logarithmic, i.e. an almost directly proportional on a logarithmic scale..

The K index is measured at many different places around the world. The magnetic field varies around the globe and accordingly a different value for K is created at each measurement station. Owing to the fact that the magnetic field varies in different ways around the globe dependent upon the way in which the magnetosphere is affected, it is not possible to have a simple relationship between one station and a global K index. Instead the individual K indices are averaged around the globe to give what is termed the Kp or planetary K index.

Kp Index: The planetary or Kp index has values that range between 0 and 9. The values of the index give a good indication of geomagnetic activity: values between 0 and 1 indicate quiet magnetic conditions and would give rise to virtually no degradation in HF band radio communications conditions. Values between 2 and 4 provide an indication of unsettled magnetic conditions that indicate the possibility of some degradation on the HF bands for radio communications. A value of 5 signifies a minor storm and 6 a larger one. Values through to 9 indicate steadily worsening conditions with 9 representing a major storm that is likely to result in a blackout in HF ionospheric propagation for several hours.

A Index; The A index is a linear measure of the Earth's field. As a result of this, its values extend over a much wider range. It is derived from the K index by scaling it to give a linear value which is termed the "a" index. This is then averaged over the period of a day to give the A index. Like the K index, values are averaged around the globe to give the planetary Ap index.

Values for the A index range up to 100 during a storm and may rise as far as 400 in a severe geomagnetic storm.

Relationship between "K" and "a" Indices

Ap Index	Kp Index	Description
0	0	Quiet
4	1	Quiet
7	2	Unsettled
15	3	Unsettled
27	4	Active
48	5	Minor storm

Ap Index	Kp Index	Description
80	6	Major storm
132	7	Severe storm
208	8	Very major storm
400	9	Very major storm

Geomagnetic and ionospheric storms are very closely related. However they are separate effects. Geomagnetic storms relate to disturbances of the Earth's magnetic field, and ionospheric storms relate to disturbances of the ionosphere. However it is found that geomagnetic storms often lead to ionospheric ones, but not on every occasion.

Interpreting the figures

The easiest way to use these figures for radio propagation prediction is to enter them into radio propagation prediction software. This will provide the most accurate prediction of what might be happening. These programmes will take into account factors such as signal paths because some will cross the poles and they will be far more affected by storms that will those across the equator.

However it is still possible to gain a good insight into what the figures mean in terms of radio propagation for all forms of radio communications using ionospheric propagation purely by assessing them mentally. Obviously high levels of solar flux are needed for good radio communications propagation. Generally the higher the flux the better the conditions will be. However the levels need to be maintained for some days. In this way the overall level of ionisation in the F2 layer will build up. Typically values of 150 and more will ensure good HF propagation conditions, although levels of 200 and more will ensure they are at their peak. In this way the maximum useable frequencies will rise, thereby providing good conditions for HF band radio communications.

The level of geomagnetic activity has an adverse affect, depressing the maximum useable frequencies. The higher the level of activity and hence the higher the Ap and Kp indices the greater the depression of the MUFs. The actual amount of depression will depend not only on the severity of the storm, but also its duration.

Summary

Having an understanding of the solar indices is of great help when using HF radio communications, be it two way radio communications, mobile radio communications, radio broadcasting or any form of point to point radio communications using ionospheric or HF propagation. It helps with radio propagation prediction and enables a quick assessment to be made of the possibility of communications being disrupted. Also having a general understanding enables any errors in entering data into the programmes to be quickly noted and corrected. In this way it enables the best to be made of the radio communications equipment and the most reliable communications to be obtained.

Radio Propagation Prediction Software

- an overview, article, tutorial about the basics of radio propagation prediction software for use on the HF bands for predicting broadcast coverage, as well as the possibility of establishing two way radio communications, maritime mobile radio communications, and for many more applications.

This radio path loss and link budget tutorial is split into several pages each of which address different aspects of radio path loss and link budget:

[1] HF ionospheric propagation basics [2] The ionosphere [3] Critical frequency, MUF and LUF [4] HF ionospheric propagation and sunspots [5] Ionospheric absorption [6] How the sun affects radio signal propagation [7] Sunspots [8] Solar disturbances [9] SID sudden ionospheric disturbance [10] Solar indices for prediction radio propagation [11] Radio propagation prediction software [12] NVIS near vertical incidence skywave [13] Sporadic E

As electromagnetic waves, and in this case, radio signals travel, they interact with objects and the media in which they travel. As they do this the radio signals can be reflected, refracted or diffracted. These interactions cause the radio signals to change direction, and to reach areas which would not be possible if the radio signals travelled in a direct line.

While it is possible to gain an idea of what HF radio propagation conditions may be like, and what the likelihood of establishing a particular radio communications link may be, purely from a general knowledge of HF radio propagation and a knowledge of the solar indices. However for a more exact analysis of the possibility of establishing two way radio communications, mobile communications, or for assessing broadcast coverage it is necessary to undertake a more rigorous and mathematical approach. This lends itself very well to a computer and software orientated approach.

There are very many radio propagation prediction software packages that are available. Many of these radio propagation prediction software packages have been developed over many years. Some were developed for broadcasters, while others were developed for military applications. With the reduction in the use of the HF bands for military and broadcast applications many of these software packages have been released for free use or for purchase at a minimal cost. This has considerably widened the use of these packages which are now within the reach of many radio communications users.

This radio propagation prediction software can now be freely used by all manner of radio users including two way radio communications applications, maritime mobile radio communications, radio broadcasters (small and large), point to point radio applications, radio amateurs, etc. In this way all these users are able to choose the optimum frequencies and times of day to establish their radio communications link, or to broadcast to a particular area with a far greater degree of certainty of being able to successfully achieve their goal.

Radio propagation prediction software

Radio propagation prediction software uses many factors to determine the statistical probability of a particular path being open. Built in to the programme there is always a considerable amount of data that has been gained over many years of research. Items such as the time of day, the variations of HF propagation that are experienced over different areas of the globe. The season and time of day also have significant impacts of the radio propagation prediction as well..

In addition to these various other inputs are required to enable the radio propagation predictions to be made. The indicators of the level of solar radiation and geomagnetic activity are also needed. These may be required in various forms but they can generally be obtained from a variety of sources on the Internet. These need to be entered along with the required path or coverage area needed and obviously the location of the transmitter.

Radio propagation prediction software packages

There is a large variety of radio propagation software packages that have been developed. Many are very sophisticated and have been developed over many years. These radio propagation software packages are used by all users of the HF radio communications from broadcasters who need to know the likely coverage areas for their radio transmissions and other professionals radio communications users who need to be able to know which frequencies to use to establish a link to a given area to radio amateurs.

NVIS Propagation

- overview, article, tutorial about NVIS, Near Vertical Incidence Skywave propagation used for radio communications and radio broadcasting.

This radio path loss and link budget tutorial is split into several pages each of which address different aspects of radio path loss and link budget:

[1] HF ionospheric propagation basics [2] The ionosphere [3] Critical frequency, MUF and LUF [4] HF ionospheric propagation and sunspots [5] Ionospheric absorption [6] How the sun affects radio signal propagation [7] Sunspots [8] Solar disturbances [9] SID sudden ionospheric disturbance [10] Solar indices for prediction radio propagation [11] Radio propagation prediction software [12] NVIS near vertical incidence skywave [13] Sporadic E

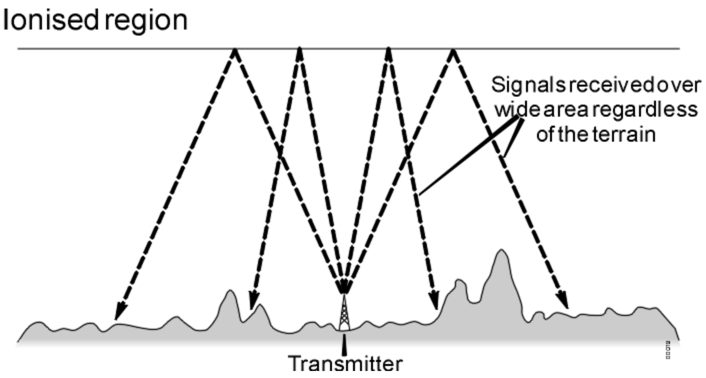
As electromagnetic waves, and in this case, radio signals travel, they interact with objects and the media in which they travel. As they do this the radio signals can be reflected, refracted or diffracted. These interactions cause the radio signals to change direction, and to reach areas which would not be possible if the radio signals travelled in a direct line.

NVIS, or Near Vertical Incidence Skywave propagation is a form of ionospheric radio propagation that can be used where radio communications links are needed to cover short distances. NVIS is particularly useful where radio communications coverage is required in regions where the ground is mountainous or rough because other modes relying on more direct coverage have significant areas where the radio signal is masked or shadowed. If MF or HF radio communications were used, the ground wave would be attenuated by the hills and mountains, and similarly line of sight radio communications using VHF or above would not be possible. Using Near Vertical Incidence Sky Wave (NVIS) it is possible to have a signal with a high angle of elevation that is not shielded by the terrain.

In view of this, NVIS is particularly suited to professional radio communications applications in geographical regions that are poorly served by land lines or line of sight repeaters, and it can provide a very useful and cost effective alternative to satellite links. It also has advantages over VHF radio in many applications for built up or forested regions as the buildings and trees introduce very significant levels of loss making line of sight VHF radio communications links virtually unworkable in many instances.

Use of NVIS

Using NVIS propagation, a high angle or near vertical signal is transmitted towards the ionosphere. This must be at a frequency that is below the critical frequency, i.e. the maximum frequency at which a vertically incident signal is "reflected" by the ionosphere. The near vertical incident signal is "reflected" by the ionosphere and returned to the Earth over an area of many kilometres either side of the transmitter. In this way good local coverage can be obtained. This form of propagation is particularly useful in rough terrain because the coverage area is illuminated from above and undulations in the Earth's surface do not create shadow areas that cannot be reached.



Near Vertical Incidence Skywave, NVIS

Typically radiation should be at angles greater than 75 or 80 degrees to the horizontal so that good local coverage is achieved. Typically coverage areas may have a radius of between 35 and 350 km. The frequency needs to be chosen carefully. Usually this is between about 2 and 10 MHz, although during the periods of the sunspot minimum, maximum frequencies may be limited to 6 to 7 MHz. By using these frequencies, the losses from the D region can be overcome, and the higher layers of the ionosphere are still able to reflect the high angle signals without them passing right through the ionosphere.

Antennas for NVIS

In order to be able to make use of NVIS for radio communications systems, it is necessary to radiate the majority of power at a high angle of radiation, i.e. near to the vertical. This requires the antennas to be designed specifically with this application in mind. In many instances HF radio antennas are designed to provide a much lower angle of radiation. This requires radio antennas specifically designed to enable the required high angle of radiation to be achieved. Typically this is achieved by using a horizontally polarised antenna. In addition to this the antenna should be mounted at an elevation that will maximise the high angle of radiation. Normally this means that it will not be particularly high.

NVIS summary

Although NVIS is has not been particularly widely used as a concept until recently, its use is growing, especially for specialised applications such as broadcasting in what are termed the "Tropical" broadcast bands where it is possible to achieve a relatively local coverage area, especially in mountainous or forested regions. In addition to this, NVIS can also be used for a number specialist two way radio communications or mobile radio communications applications.

Sporadic E

This radio path loss and link budget tutorial is split into several pages each of which address different aspects of radio path loss and link budget:

[1] HF ionospheric propagation basics [2] The ionosphere [3] Critical frequency, MUF and LUF [4] HF ionospheric propagation and sunspots [5] Ionospheric absorption [6] How the sun affects radio signal propagation [7] Sunspots [8] Solar disturbances [9] SID sudden ionospheric disturbance [10] Solar indices for prediction radio propagation [11] Radio propagation prediction software [12] NVIS near vertical incidence skywave [13] Sporadic E

As electromagnetic waves, and in this case, radio signals travel, they interact with objects and the media in which they travel. As they do this the radio signals can be reflected, refracted or diffracted. These interactions cause the radio signals to change direction, and to reach areas which would not be possible if the radio signals travelled in a direct line.

Sporadic E is a form of propagation that can arise with little warning, and enable frequencies of 100 MHz and more to travel over distances of a thousand kilometres and more. Many people experienced it in the days of the old VHF television transmissions. When sporadic E propagation arose, it would result in severe interference to the signals. Even now VHF FM broadcasts in the 88 - 108 MHz band can be affected. In many instances the arrival of sporadic E can cause unwanted interference as signals that are normally too distant to be heard appear. As a result it is of great interest when planning a radio system or network and for predicting achievable distances..

What is sporadic E?

Sporadic E arises when clouds of intense ionisation form in the region of the E layer. These clouds can have very high levels of ionisation, allowing frequencies up to about 150 MHz to be reflected on some occasions. The clouds are usually comparatively small, measuring only about 50 to 150 kilometres in diameter. Their shape is irregular. Sometimes they may be almost circular, whereas others may be long and thin. They are also surprisingly thin, often only measuring a few hundred metres in depth.

The clouds appear almost at random, although there are times when they are more likely to occur. They form in the day, and dissipate within a few hours. They are also far more common in summer, peaking approximately in mid summer. As they form the level of ionisation gradually builds up, affecting first the lower frequencies, and later higher frequencies as the level of ionisation increases.

Propagation via sporadic E occurs in the same way as normal ionospheric propagation. Signals from the transmitter leave the earth as a sky-wave, travelling towards the ionosphere. Here they are reflected (or more correctly refracted) back to earth where they are heard at a considerable distance from the transmitter. Like normal ionospheric propagation it is the free electrons that affect the signals, causing them to bend back towards the earth. In view of the fact that the sporadic E clouds occur at around the same height as the E layer, similar distances are achieved. Typically the maximum distances are about 2000 km.

It is found that the sporadic E ionisation clouds move. Being in the upper atmosphere they are blown by the winds in these areas and can drift at speeds of up to 300 kilometres per hour. This means that when sporadic E is being experienced, the area from which stations are heard will change over the life of the cloud.

Sporadic-E theories

There are many theories about sporadic E and how it occurs. Some believe that it may be related to thunderstorms, others think it results from the winds in the upper atmosphere. None of these theories have been established, leaving the reasons behind sporadic E a mystery, and predictions of when it will occur have to be left to statistics. However even though the mechanism behind the formation of sporadic E is not fully known it is still possible for radio amateurs to utilise them to enable them to make contacts over long distances.

Signal propagation for satellites

- the effects of the atmosphere on satellite signals

Satellites are widely used these days for everything from navigation, in the case of GPS, satellite television broadcasting, communications, mobile phone technology, Internet broadband weather monitoring and much more. They normally use frequencies that are in excess of 500 MHz where the signals are not unduly affected by the ionosphere or troposphere. However some effects can be noticed and are important, especially when planning, installing or setting up a satellite system.

Ground to satellite paths

When signals travel from the ground up to the satellite they pass through four main regions. These are the troposphere, above which is region that is often termed inner free space which is above the troposphere and below the ionosphere. The next region is the ionosphere, and finally there is the outer free space.

There are a number of different of effects that are introduced by these regions. Transmission in free space has unity refractive index and is loss-less (apart from the spreading effect that reduces the signal power over a fixed area with distance away from the source, but no power is actually lost).

The troposphere and ionosphere have refractive indices that differ from unity. The troposphere is greater than unity and the ionosphere is less than unity and as a result refraction and absorption occur. The inner free space region also has little effect.

Faraday rotation

A further effect that is introduced by the ionosphere is known as Faraday rotation which results from the fact that the ionosphere is a magneto-ionic region. The Faraday rotation of a signal causes different elements of a signal to travel in different ways, particularly rotating the plane of polarisation. This can create some problems with reception. A linearly polarised signal can be considered as two contra-rotating circularly polarised signals. The phase velocities of these two signals vary in a magnetic medium such as the ionosphere and as a result the polarisation of the signal changes. The degree of change is dependent upon the state of the ionosphere and it follows the same pattern as that experienced for HF ionospheric communications changing over the course of the day, with the seasons and over the sunspot cycle.

Ionospheric scintillations

Another of the effects introduced by the ionosphere is termed "ionospheric scintillations." These scintillations manifest themselves as a variety of variations of amplitude, phase, and polarisation angle. They can also change the angle of arrival of the signals. These variations change over a period of between one to fifteen seconds, and they can affect signals well into the microwave region.

The variations are caused primarily by the variations in electron density arising in the E region, often as a result of sporadic E but also in the F layer where a spreading effect is the cause. The level of scintillation is dependent upon a number of factors including the location of the earth station and the state of the ionosphere, as a result of the location, the sunspot cycle, the level of geomagnetic activity, latitude, and local time of day.

The scintillations are more intense in equatorial regions, falling with increasing latitude away from the equator but then rising at high latitudes, i.e. in the auroral zone or the region where auroras take place. The effects are also found to decrease with increasing frequency, and generally not noticeable above frequencies of 1 - 2 GHz. As such they are not applicable to many direct broadcast television signals, although they may affect GPS, and some communications satellites.

Tropospheric effects

There are a number of effects that the troposphere introduces including signal bending as a result of refraction, scintillation, and attenuation.

The signal refraction in the troposphere is in the opposite sense to that in the ionosphere. This is because the refractive index in the troposphere is greater than unity, and it is also frequency independent. The signal refraction gives them a greater range than would be expected as a result of the direct geometric line of sight. Tropospheric ducting and extended range effects that are experienced by terrestrial VHF and UHF communications may also be experienced when low angles of elevation are used.

Scintillations induced by the troposphere are often greater than those seen as a result of the ionosphere. They occur as a result of the turbulence in the atmosphere where areas of differing refractive index move around as a result of the wind or convection currents. The degree to which the scintillations occur is dependent upon the angle of inclination, and above angles of around 15 degrees the effect can normally be ignored. At angles between 5 and 10 degrees the changes can often be around 6 dB at frequencies of around 5 GHz.

Doppler shift

Frequency changes as a result of the Doppler shift principle may be in evidence with signals from some satellites. Satellites in Low Earth Orbits move very quickly, and as a result a Doppler frequency shift is apparent in many cases. With the satellite moving towards the earth station the frequency appears higher than nominal, and then as it moves away the apparent frequency falls. The degree of shift is dependent upon a number of factors including the speed of the satellite (more correctly its speed relative to the earth station) and the frequencies in use. Shifts of the order of 10 kHz may be experienced. As most satellites operate in a cross mode configuration, the Doppler shift is not just applicable to the band on which the signal is received, but to the cumulative effect of the uplink and downlink transmissions. In many instances the effects will subtract because of the way the satellite mixing process is configured.

Summary

Although satellites generally operate at frequencies that may be thought to be immune from tropospheric and ionospheric disturbance, these regions still have a significant effect and this needs to be taken into account when designing satellite systems.

Radiowave Propagation and the Atmosphere

- overview of radio signals and radiowave propagation and how different areas of the atmosphere affect radio communications.

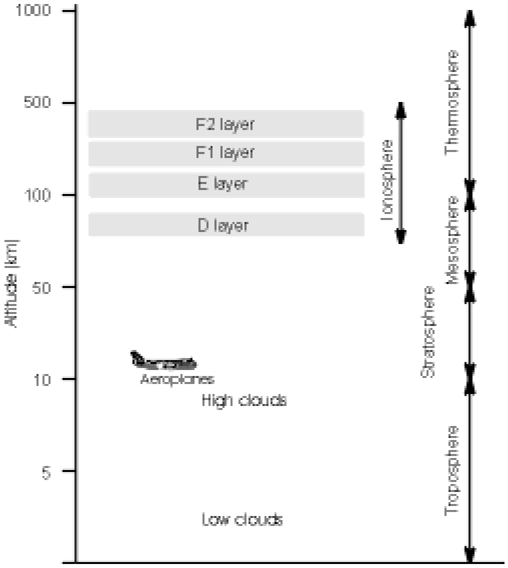
The way that radio signals propagate, or travel from the radio transmitter to the radio receiver is of great importance when planning a radio communications network or system. This is governed to a great degree by the regions of the atmosphere through which they pass. Without the action of the atmosphere it would not be possible for radio communications signals to travel around the globe on the short wave bands, or travel greater than only the line of sight distance at higher frequencies. In fact the way in which the atmosphere affects radio communications is of tremendous importance for anyone associated with radio communications, whether they are for two way radio communications links, mobile radio communications, radio broadcasting, point to point radio communications or any other radio.

In view of the importance of the atmosphere to radio communications, an overview of its make-up is given here.

Layers of the Atmosphere

The atmosphere can be split up into a variety of different layers according to their properties. As different aspects of science look at different properties there is no single nomenclature for the layers. The system that is most widely used is that associated with. Lowest is the troposphere that extends to a height of 10 km. Above this at altitudes between 10 and 50 km is found the stratosphere. This contains the ozone layer at a height of around 20 km. Above the stratosphere, there is the mesosphere extending from an altitude of 50 km to 80 km, and above this is the thermosphere where temperatures rise dramatically.

There are two main layers that are of interest from a radio communications viewpoint. The first is the troposphere that tends to affect radio frequencies above 30 MHz. The second is the ionosphere. This is a region which crosses over the boundaries of the meteorological layers and extends from around 60 km up to 700 km. Here the air becomes ionised, producing ions and free electrons. The free electrons affect radio communications and radio signals at certain frequencies, typically those radio frequencies below 30 MHz, often bending them back to Earth so that they can be heard over vast distances around the world.



Regions of the Atmosphere

Troposphere

The lowest of the layers of the atmosphere is the troposphere. This extends from ground level to an altitude of 10 km. It is within this region that the effects that govern our weather occur. To give an idea of the altitudes involved it is found that low clouds occur at altitudes of up to 2 km whereas medium level clouds extend to about 4 km. The highest clouds are found at altitudes up to 10 km whereas modern jet airliners fly above this at altitudes of up to 15 km.

Within the troposphere there is generally a steady fall in temperature with height and this has a distinct bearing on some radio propagation modes and radio communications that occur in this region. The fall in temperature continues in the troposphere until the tropopause is reached. This is the area where the temperature gradient levels out and then the temperature starts to rise. At this point the temperature is around -50 �C.

The refractive index of the air in the troposphere plays a dominant role in radio signal propagation and the radio communications applications that use tropospheric radiowave propagation. This depends on the temperature, pressure and humidity. When radio communications signals are affected this often occurs at altitudes up to 2 km.

The ionosphere

The ionosphere is an area where there is a very high level of free electrons and ions. It is found that the free electrons affect radio waves and hence they have a marked effect on radio communications in many instances. Although there are low levels of ions and electrons at all altitudes, the number starts to rise noticeably at an altitude of around 30 km. However it is not until an altitude of approximately 60 km is reached that the it rises to a sufficient degree to have a major effect on radio signals.

The overall way in which the ionosphere is very complicated. It involves radiation from the sun striking the molecules in the upper atmosphere. This radiation is sufficiently intense that when it strikes the gas molecules some electrons are given sufficient energy to leave the molecular structure. This leaves a molecule with a deficit of one electron that is called an ion, and a free electron. As might be expected the most common molecules to be ionised are nitrogen and oxygen.

Most of the ionisation is caused by radiation in the form of ultraviolet light. At very high altitudes the gases are very thin and only low levels of ionisation are created. As the radiation penetrates further into the atmosphere the density of the gases increases and accordingly the numbers of molecules being ionised increase. However when molecules are ionised the energy in the radiation is reduced, and even though the gas density is higher at lower altitudes the degree of ionisation becomes less because of the reduction of the level of ultraviolet light.

At the lower levels of the ionosphere where the intensity of the ultraviolet light has been reduced most of the ionisation is caused by x-rays and cosmic rays which are able to penetrate further into the atmosphere. In this way an area of maximum radiation exists with the level of ionisation falling below and above it.

In terms of its radio communications properties, the ionosphere is often thought of as a number of distinct layers. Whilst it is very convenient to think of the layers as separate, in reality this is not quite true. Each layer overlaps the others with the whole of the ionosphere having some level of ionisation. The layers are best thought of as peaks in the level of ionisation. These layers are given designations D, E, and F1 and F2.

Description of the layers in the ionosphere

1. ***D layer:*** The D layer is the lowest of the layers of the ionosphere. It exists at altitudes around 60 to 90 km. It is present during the day when radiation is received from the sun. However the density of the air at this altitude means that ions and electrons recombine relatively quickly. This means that after sunset, electron levels fall and the layer effectively disappears. This layer is typically produced as the result of X-ray and cosmic ray ionisation. It is found that this layer tends to attenuate signals that pass through it.

2. ***E layer:*** The next layer beyond the D layer is called the E layer. This exists at an altitude of between 100 and 125 km. Instead of acting chiefly as an attenuator, this layer reflects radio signals although they still undergo some attenuation.

In view of its altitude and the density of the air, electrons and positive ions recombine relatively quickly. This occurs at a rate of about four times that of the F layers that are higher up where the air is less dense. This means that after nightfall the layer virtually disappears although there is still some residual ionisation, especially in the years around the sunspot maximum that will be discussed later.

There are a number of methods by which the ionisation in this layer is generated. It depends on factors including the altitude within the layer, the state of the sun, and the latitude. However X-rays and ultraviolet produce a large amount of the ionisation light, especially that with very short wavelengths.

3. ***F layer:*** The F layer is the most important region for long distance HF communications. During the day it splits into two separate layers. These are called the F_1 and F_2 layers, the F_1 layer being the lower of the two. At night these two layers merge to give one layer called the F layer. The altitudes of the layers vary considerably with the time of day, season and the state of the sun. Typically in summer the F_1 layer may be around 300 km with the F_2 layer at about 400 km or even higher. In winter these figures may be reduced to about 300 km and 200 km. Then at night the F layer is generally around 250 to 300 km. Like the D and E layers, the level of ionisation falls at night, but in view of the much lower air density, the ions and electrons combine much more slowly and the F layer decays much less. Accordingly it is able to support radio communications, although changes are experienced because of the lessening of the ionisation levels. The figures for the altitude of the F layers are far more variable than those for the lower layers. They change greatly with the time of day, the season and the state of the Sun. As a result the figures which are given must only be taken as an approximate guide.

   Most of the ionisation in this region of the ionosphere is caused by ultraviolet light, both in the middle of the UV spectrum and those portions with very short wavelengths.

Summary

The way in which the various regions in the atmosphere affect radiowave propagation and radio communications is a fascinating study. There are very many factors that influence radio propagation and the resulting radio communications links that can be established. Predicting the ways in which this occurs is complicated and difficult, however it is possible to gain a good idea of the likely radio communications conditions using some simple indicators. Further pages in this section of the website detail many of these aspects.

Practical Applications for Distributed Antenna Systems

Simon Jones of AIB Wireless looks at the way Distributed Antenna Systems, DAS, can be used to provide widespread wireless coverage

With an ever increasing requirement for high bandwidth mobile broadband on cellular networks, a Distributed Antenna System (DAS) can provide an attractive solution. One such system on Oxford Street is owned by shared infrastructure provider Wireless Infrastructure Group (WIG). This system has the potential to provide wireless coverage for multiple operators and multiple technologies for 2km along this busy urban location.

Radio System Design

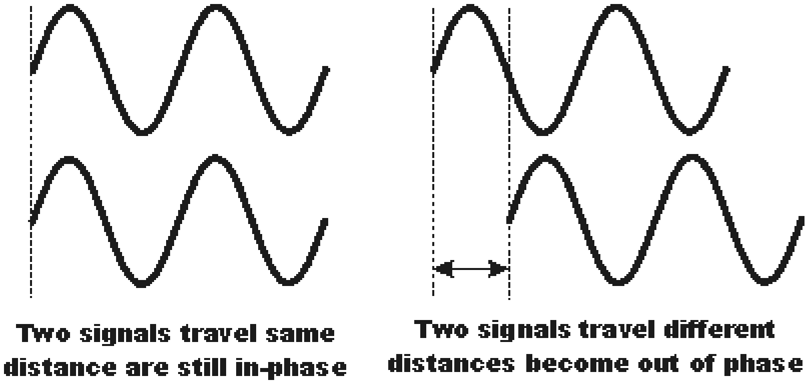
Radio waves are all around us and are used to carry prolific amounts of information to all points of the globe. Even the most modest of transmitting equipment can produce a signal that can effectively be received over millions of cubic meters of space. Radio waves travel at the speed of light and can cover huge distances with little loss of intensity. Radio telescopes routinely detect signals that have travelled billions of miles over millions of years. They do have one drawback however. Unlike other forms of radiation they can be stopped dead in there tracks by something as insignificant as a piece of tin foil. Other materials are more or less 'transparent' to radio signals than others. Air for example has very little effect on a radio transmission, stone walls do.

Because of this 'shadowing' effect the design of a radio system covering a building or dense urban environment is very different to a similar system covering a rural or less cluttered outdoor area.

Fading

As well as attenuation of the radio signal as is passes through walls and floors or any other relatively dense structure, another problem occurs that is generally referred to as 'fading'. This effect is seen in any wave like transmission - even light waves. The result of fading is a large variation in signal strength over a short distance. Fading occurs when a signal arrives at a point simultaneously from 2 directions. E.g. 1 signal arrives directly; the other one has bounced off the side of a building and also arrives at the same point. Because the signals have effectively travelled different distances their wave forms may no longer coincide with each other and have a difference in 'phase'.

The effect of this is to cause partial or complete cancellation of the two radio signals creating a large variation in the signal level detected by the receiver

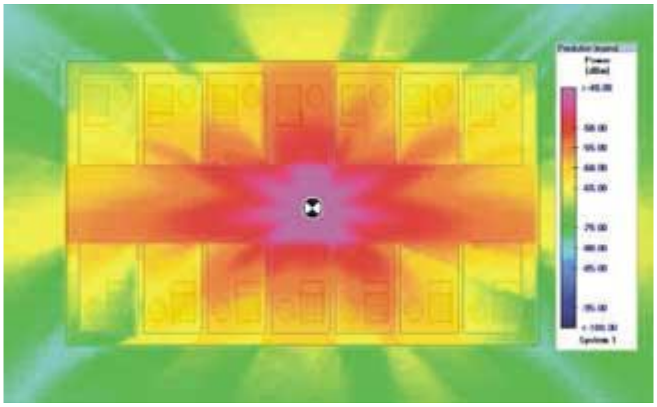

Two signals travel same distance are still in-phase
Two signals travel different distances become out of phase

Signal phasing from different antennas

The above illustration shows firstly two signals in phase. The effect of this is a slight increase in the received level. The second illustration shows 2 signals perfectly out of phase. If both signals are of the same amplitude, the signal can theoretically be completely eliminated. Even a significant increase in the power of the transmitter would make no difference.

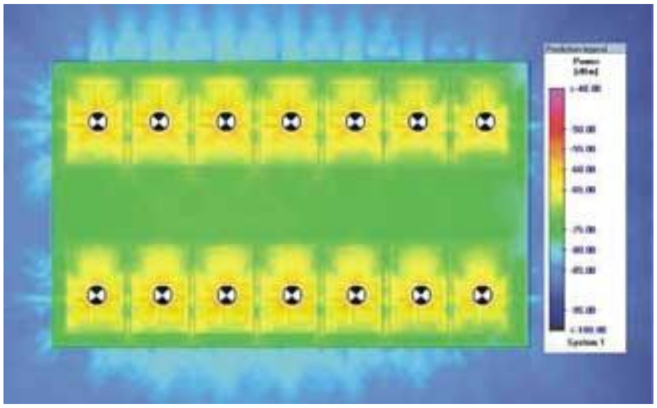
Coverage Concepts

To overcome coverage problems the simple answer would be to have a higher power transmitter. However within buildings and dense urban areas the opposite is true. Radiating from a single location exacerbates the effects of shadowing and fading described above. The following illustrations show computer generated coverage predictions for a single high power point source system along with a system using several low power point sources.



Coverage predictions for a single high power point source

Single point source shows good coverage near the antenna but varying coverage around the building with much of the signal being lost outside.



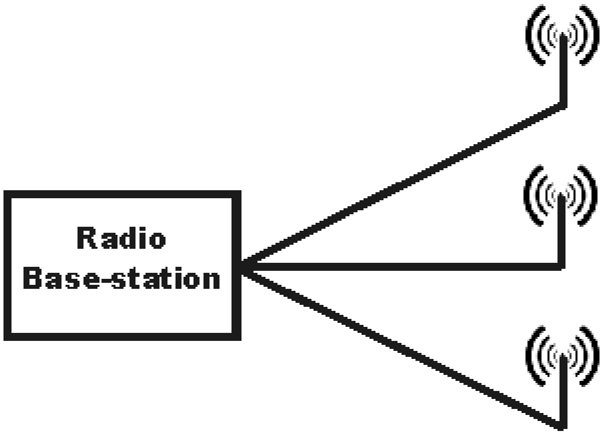
Coverage predictions using multiple low power sources

Multiple point low level point sources result in a more even signal distribution with less radiation outside the building.

Distributed Antenna System DAS Basics

For building and dense urban environments therefore, a system is required to distribute the radio signal more evenly from several point sources rather than a single one, hence the concept of a Distributed Antenna System.

The concept of a conventional distributed antenna system is shown below:



Multiple antennas used by a Base-Station

In simple terms most Distributed Antenna Systems are made up of a central base station connected to several antennas via conventional copper coaxial cable. This system would be described as 'Passive' in that all of the distribution components such as the cable itself along with any splitters, combiners, couplers attenuators or other components do not require additional power to function.

The high power signal from the base station is distributed along the coaxial cable usually to a collection of 'Nodes' where a small amount of the signal is 'tapped off' to feed individual sections of the DAS. The technique for this is generally referred to as 'Trunk and Spur' due to its tree like topology.

The advantages to this approach are its simplicity and high reliability. In most cases nothing short of actual physical damage to a system component will cause a failure.

The system does, however suffer from one major drawback common to any system employing copper based distribution, resistance!

Like any other form of electrical signal, radio transmissions use electrical currents to travel from the base station to the transmitting antenna and, like any form of electrical current they can be affected by the finite resistance of the copper based cables used to distribute them. The effect of this is a gradual loss of the signal level along the cable. This results in a finite distance the radio signal transmission can travel before it is completely dissipated as heat along the cable. This effect is worse at higher frequencies and is also subject to a phenomenon called 'Skin Effect' where the signal does not travel over the whole thickness of the copper conductor. Other factors also cause loss of the signal along the cable such as partial radiation of the valuable RF signal.

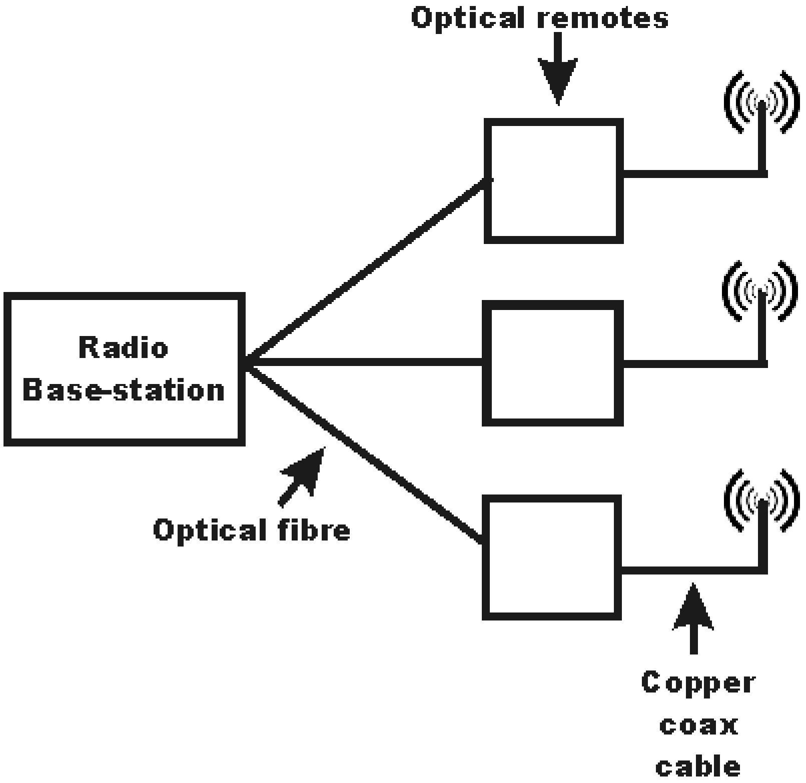
The obvious way around this is to make the cable thicker so reducing the electrical resistance. This has practical limitations such as the cost of the cable and the actual process of installing thick cables within building infrastructure or along streets. A copper based DAS is therefore limited in the area it can cover particularly at higher frequencies. A VHF system at 160MHz could potentially cover cable runs of several hundred meters whereas a cellular 3G 2100 MHz system would struggle over more than 200 metres. Higher frequency systems also require more radiating antennas as their coverage properties through the air are reduced due to the increased wavelength and reduced antenna field strength.

Where long distances are involved between the base station equipment and radiating antennas, an alternative distribution technology is required.

RF over fibre

The technology deployed by the Oxford Street DAS is Radio Frequency over Fibre (RFoF). Fibre optic systems have existed for many years and are often used to distribute data over long distances. Signals are converted to optical (light) waves generally using a laser diode. These light waves are injected in to a single piece of very thin glass fibre which is used to connect to a similar apparatus at the remote end. Fibre optic signals not only offer wide band width due to the extremely high frequency of light waves but are also subject to very little loss along the fibre optic cable. A typical optical system can cover more than 20 km while only suffering from a 0.5 dB (10%) loss in signal per km. Compare this to 100 dB (99.99999999%)loss for a high quality coaxial cable at 1000 MHz and you can begin to see the advantages.

So how are optical system used to distribute radio waves? - Clearly some kind of conversion process needs to take place before a radio system can be connected to a fibre optic cable. This is achieved by connecting the radio base station to an optical 'master' unit. This interfaces to the radio equipment and performs the essential conversion process. Fibre optic cables from the master unit are then used to connect to a series of optical 'remote' units. These powered units convert the optical signal back to radio frequencies which are then distributed to radiating antennas using a conventional copper based network. Although copper is still used at each end, the optical system eliminates the requirement for long runs of copper coaxial cable.



DAS using optical remote stations

So, master and remote units carry out the conversion to and from an optical signal in both directions (transmit and receive) with the rest of the DAS consisting of standard passive components. This solves the problem of distance distribution but does have disadvantages of its own.

Additional equipment means additional cost. All of the optical equipment is active and generally complex so is gives rise to reliability and maintenance issues. The remote units also require electrical power at each location which gives rise to the necessity for a separate 240volt power distribution system.

The remote units are also generally limited in the range of frequencies they can accommodate and the amount of power they can generate. Unlike passive systems which for example can generally cover all 3 cellular bands, normally one remote unit per band would be required.

Modern cellular systems need to cover all 3 operating bands so this could in effect mean 3 remote units at each location. On complex cellular systems using many individual transmitters (carriers) the coverage from each remote unit can be severely limited due to the power available for each of these carriers.

Hybrid systems

The inclusion of optical components on a typical DAS often gives rise to a Hybrid system. This type of deployment uses antennas connected directly to the base station via copper as well as antennas utilising the optical system. The reason for this is that most base station equipment is intended for external use and hence has a relatively high output power of 25 to 50 Watts per channel. Optical systems however require very little signal level to drive them, typically 1 mW so there is a significant excess of power to be used conventionally. If all of the antennas are located some distance from the base station then the excess power is often dissipated as heat in a dummy load. This is a waste of energy and in the current energy conservation era in which we live, is a practice becoming ever more frowned upon.

Combining equipment

Another benefit of using optical equipment is the reduced cost of the combining equipment for multiple technologies. With conventional systems where multiple technologies and services are connected, a specialised combiner is required to avoid interaction between systems connected to the DAS. This type of equipment generally takes the form of multiple high quality filtering devices connected to the common DAS output. Isolation requirements for this equipment can be very high particularly for cellular systems which results in an expensive device which can rival the cost of the rest of the DAS. Because the optical system uses very low power levels, the required isolation can be achieved by the simple attenuation devices used to drop the drive power to the required levels. This represents a significant reduction in cost compared to the standard high power combiner.

Unfortunately the Hybrid system loses out as it still requires the high power combiner equipment as well as components to drive the optical system.

Oxford Street DAS

The DAS at Oxford Street in London, UK was originally designed to provide 2G coverage from seven antennas along Oxford Street. Each antenna is connected back to cellular base station equipment located in the basement of the Marble Arch Thistle Hotel.

This particular DAS is not a hybrid design due to the distances between base station and antenna and also because of the installation challenges involved in connecting street level antennas to the base station equipment.

Each remote unit is located within existing CCTV equipment cabinet and connected to a local antenna via conventional copper. The antennas are located on top of street lamp poles at the major road intersections.

The original system design was based on a 2G deployment, following a planned upgrade the system will support 3G and WiFi.

Urban DAS benefits

Being located in a dense urban environment with many local roof top cell sites one could ask the question 'why deploy a DAS in this location'? Surely there is sufficient coverage from these roof tops sites? Adequate cellular coverage would be available without the Oxford Street DAS but the system does have several advantages over a roof top 'macro' based network.

1. Signals from the base station are radiated evenly along the Oxford St DAS resulting in a continuous signal level.
2. Antennas are in close proximity to subscribers resulting in consistent signal quality and higher quality of service.
3. Penetration into nearby shops and cafes is improved compared to the 'coverage from above' of rooftop sites.
4. Traffic capacity in the area is improved and evenly distributed along the DAS In this urban environment DAS has the potential to reach more subscribers than roof top systems.
5. System capacity can be tailored for each radiating antenna to maximize efficiency.
6. The network can be efficiently and cost effectively extended as far as the fibre optic network reaches.

AIB Wireless helps organisations gain the maximum benefit - with the minimum level of risk - from wireless technologies. Our portfolio of services includes wireless surveys, audits, wireless network design and management. AIB Wireless assists WIG in the maintenance of the Oxford St DAS.

Wireless Infrastructure Group - WIG is a leading provider of shared communications infrastructure in the UK. The company owns and manages over 1,000 radio sites and is planning to deploy a number of operator neutral DAS solutions over the next five years. WIG works closely with the building, venue or local authority host to agree an acceptable coverage solution before making delivery commitments to the operators.

One of the founders of AIB Wireless, Simon Jones has more than 20 years experience in the design, development and deployment of electronic communication and control sys